Decoding with Confidence: Statistical Control on Decoder Maps

Jérôme-Alexis Chevalier, Tuan-Binh Nguyen, Joseph Salmon, Gaël Varoquaux, Bertrand Thirion

Decoding with Confidence: Statistical Control on Decoder Maps

Jérôme-Alexis Chevalier[1,2,3], Tuan-Binh Nguyen[1,2,3,4], Joseph Salmon[5], Gaël Varoquaux[1,2,3], Bertrand Thirion[1,2,3]

jerome-alexis.chevalier@inria.fr

[1] Parietal project-team, Inria Saclay-Ile de France, Palaiseau, France
[2] CEA/Neurospin bat 145, Gif-Sur-Yvette, France
[3] Université Paris-Saclay, Gif-Sur-Yvette, France
[4] LMO, Université Paris-Saclay, Orsay, France
[5] IMAG, Université de Montpellier, CNRS, Montpellier, France

Keywords: fMRI, Decoding, Statistical Methods, Multivariate Model, Inference, Statistical Control, Support Recovery, High Dimension.

Abstract

In brain imaging, decoding is widely used to infer relationships between brain and cognition, or to craft brain-imaging biomarkers of pathologies. Yet, standard decoding procedures do not come with statistical guarantees, and thus do not give confidence bounds to interpret the pattern maps that they produce. Indeed, in whole-brain decoding settings, the number of explanatory variables is much greater than the number of samples, hence classical statistical inference methodology cannot be applied. Specifically, the standard practice that consists in thresholding decoding maps is not a correct inference procedure. We contribute a new statistical-testing framework for this type of inference. To overcome the statistical inefficiency of voxel-level control, we generalize the Family Wise Error Rate (FWER) to account for a spatial tolerance δ, introducing the δ-Family Wise Error Rate (δ-FWER). Then, we present a decoding procedure that can control the δ-FWER: the Ensemble of Clustered Desparsified Lasso (EnCluDL), a procedure for multivariate statistical inference on high-dimensional structured data. We evaluate the statistical properties of EnCluDL with a thorough empirical study, along with three alternative procedures including decoder map thresholding. We show that EnCluDL exhibits the best recovery properties while ensuring the expected statistical control.

1 Introduction

Predicting behavior or diseases status from brain images is an important analytical approach for imaging neurosciences, as it provides an effective evaluation of the information carried by brain images. Machine learning tools, mostly supervised learning, are

indeed used on brain images to infer cognitive states [Haynes and Rees, 2006, Norman et al., 2006] or to perform diagnosis or prognosis [Demirci et al., 2008, Fan et al., 2008]. Brain images are obtained from MRI or PET imaging, or even EEG- or MEG-based volume-based activity reconstruction. They are used to predict a *target* outcome: binary (*e.g.,* two-condition tasks), discrete (*e.g.,* multiple-condition tasks) or continuous (*e.g.,* age). The *decoding models* used for such predictions are most often linear models, characterized by a weight map that can be represented as a brain image [Mourao-Miranda et al., 2005, Varoquaux and Thirion, 2014].

Besides the prediction accuracy achieved, this estimated weight map is crucial to assess the information captured by the model. Typically, the produced weight maps are used to identify discriminative patterns [Haxby et al., 2001, Mourao-Miranda et al., 2005, Gramfort et al., 2013] and support reverse inferences [Poldrack, 2011, Schwartz et al., 2013, Varoquaux et al., 2018], *i.e.,* conclude on the implication of brain regions in the studied process.

Unlike in standard analysis —statistical parametric mapping [Poldrack et al., 2011, chap 7]—, in decoding the feature importance is tested *conditional on other brain features*, *i.e.,* it assesses whether each feature *adds* to information conveyed by other features. Weichwald et al. [2015] highlight the fact that decoding, *i.e.,* multivariate or conditional analysis, and encoding, *i.e.,* univariate or marginal analysis, are complementary. They notably argue that taking the two perspectives is essential for causal interpretation regarding the implication of brain regions in the target outcome (see also Haufe et al. [2014]).

While decoding optimizes the prediction of a target outcome, little or nothing can be concluded about the significant features of weight maps. Indeed, those maps do not come with well-controlled statistical properties, making decoding models hard to interpret. For instance, considering linear Support Vector Machines (SVM) [Cortes and Vapnik, 1995] or linear Support Vector Regression (SVR) [Smola and Schölkopf, 2004], that are popular in neuroimaging [Pereira et al., 2009, Rizk-Jackson et al., 2011], a natural way to recover predictive regions from their weight maps is to threshold these maps (e.g. Mourao-Miranda et al. [2005], Rehme et al. [2015], Sato et al. [2013], Lee et al. [2010]). However, this approach is problematic for two reasons: there exists no clear way to choose the threshold as a function of a desired significance, and it is unclear whether such a thresholded map is still an accurate predictor of the outcome. Solutions that bypass the arbitrary threshold choice have been proposed, such as Recursive Feature Elimination (RFE) [De Martino et al., 2008], but the produced maps still lack statistical guarantees.

In this work, we show that the natural procedure that consists in thresholding standard decoders, such as SVR, is not a relevant solution. In this respect, we consider two thresholding strategies: one that keeps extreme weights, and another one that computes the threshold by performing a permutation test. Unlike RFE, these two thresholding strategies can be derived from statistical testing considerations —yet, these statistical properties are not assumption free. We also consider decoders that provide confidence intervals around the estimated weight map. As detailed in the next section, these ap-

proaches also face severe challenges in terms of statistical power and computational tractability. They have to rely on algorithmic shortcuts, approximations and hypotheses that are more or less problematic in practice.

Hence, for all methods considered, the control of false detections is only achieved within a certain theoretical framework, and given a series of assumptions that are not easily checked. It is thus fundamental to analyze their statistical behavior with an extensive empirical study. We present here a set of experiments assessing the accuracy of the error rate control and support recovery on real and semi-synthetic brain-imaging data.

Additionally, to achieve a reasonable compromise between error control and power, we introduce a new type of error control adapted to imaging problems. The proposed quantity is a generalization of the Family Wise Error Rate (FWER) [Hochberg and Tamhane, 1987] including a spatial tolerance parametrized by a distance δ. We call it δ-FWER.

In Section 2, we bring useful background, discuss the statistical guarantees that we aim at for pattern maps, and make the theoretical and practical inference challenges explicit. In Section 3 we provide a definition of the δ-FWER along with a geometrical interpretation of this quantity. We also describe several statistical inference methods producing statistical maps reflecting the significance of conditional association of brain regions with a target, while controlling the FWER or δ-FWER. Section 4 and Section 5 follow with extensive experiments on simulations and large-scale fMRI datasets that study the behavior of the benchmarked solutions regarding false positive control and recovery.

2 Context: decoding-map recovery

In this section, we first review a result due to Weichwald et al. [2015] about the complementarity of univariate and multivariate inference, then we present the statistical guarantees that we aim at for on brain-wide decoding maps, lastly we formalize the problem of statistical inference on such maps.

2.1 Complementarity of univariate and multivariate inference

Statistical inference in neuroimaging can be performed using a mass univariate modeling, *i.e.,* fitting brain activity maps from an outcome —leading to *encoding models*— or by predicting an outcome from brain maps using multivariate modeling —leading to *decoding models*. The complementarity of univariate and multivariate analyses has been demonstrated in Weichwald et al. [2015]. Specifically, they argued: "We showed that only encoding models in a stimulus-based setting support unambiguous causal statements. This result appears to imply that decoding models, despite their gaining popularity in neuroimaging, are of little value for investigating the neural causes of cognition. In the following, we argue that this is not the case. Specifically, we show that by combining encoding and decoding models, we gain insights into causal structure that are not pos-

sible by investigating each type of model individually." This statement clearly implies that inference tools are needed for multivariate analysis. The present work is thus fully dedicated to multivariate inference. We simply provide some univariate inference results for reference, given that they address different yet complementary questions.

2.2 Statistical control with spatial tolerance

In decoding, the signals from voxels are used concurrently to predict an outcome. Given that they display high correlations, trying to identify the effect of each covariate (voxel) is not possible. Precise voxel-level control may not be necessary: current brain models are rather specified at a regional scale, see *e.g.,* [Glasser et al., 2016]. Additionally, to control a statistical error, detecting a voxel adjacent to a truly predictive region is less problematic than detecting a false positive far from such a predictive region. These two facts argue in favor of incorporating a spatial tolerance in the sought statistical control, as with efforts in standard analysis [Smith and Nichols, 2009, Da Mota et al., 2014, Bowring et al., 2019]. Hence, we introduce a generalization of the Family Wise Error Rate (FWER) [Hochberg and Tamhane, 1987]: the δ-FWER. This generalization is related to the extension of the False Discovery Rate (FDR) [Benjamini and Hochberg, 1995] proposed by Nguyen et al. [2019] and Gimenez and Zou [2019], called δ-FDR and local-FDR, respectively.

2.3 Formal problem setting

Notation. For clarity, we use bold lowercase for vectors and bold uppercase for matrices. For $p \in \mathbb{N}$, we write $[p]$ for the set $\{1, \dots, p\}$. For a vector $\mathbf{w}$, $\mathbf{w}_j$ refers to its j-th coordinate. For a matrix $\mathbf{X}$, $\mathbf{X}_{i,j}$ refers to the element in the i-th row and j-th column.

Formalizing the decoding problem. The target (outcome to decode) is observed in n samples and denoted by $\mathbf{y} \in \mathbb{R}^n$ ($\mathbf{y}$ can be binary, discrete or continuous). The brain volume is discretized into p voxels. The corresponding p voxel signals are also referred to as explanatory variables, covariates or features. We denote by $\mathbf{X} \in \mathbb{R}^{n \times p}$ the matrix containing (column-wise) the p covariates $\{\mathbf{X}_1, \dots, \mathbf{X}_p\}$. We assume that, for all $i \in [n]$, the samples $(\mathbf{y}_i, \mathbf{X}_{i,\cdot})$ are i.i.d. Then, further assuming a linear dependency between the covariates and the response, the generative model is as follows:

$$\mathbf{y} = \mathbf{X}\mathbf{w}^* + \boldsymbol{\varepsilon} \ , \tag{1}$$

where $\mathbf{w}^* \in \mathbb{R}^p$ is the true weight map and $\boldsymbol{\varepsilon}$ is the noise vector. In the present study, we assume for simplicity that the noise is Gaussian, *i.e.,* $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_n)$, but extension to sub-Gaussian noise is possible.

High dimensionality and structure of the data. Given $\mathbf{X}$ and $\mathbf{y}$, a standard procedure computes an estimate $\hat{\mathbf{w}}$ of $\mathbf{w}^*$. Getting statistical guarantees on $\mathbf{w}^*_j$, $j \in [p]$, means assessing with some degree of uncertainty that $\mathbf{w}^*_j$ is non-zero, or equivalently,

giving a confidence interval for $\mathbf{w}_j^*$. This is hard in high dimension and when short- and long-range correlations are present in the data. Indeed, for brain imaging data, n is typically hundreds (or less), whereas p may amount to hundreds of thousands. In addition, voxel signals are highly correlated, which makes model identification harder due to multicollinearity and ill-posedness. Theoretical studies, *e.g.,* Wainwright [2009], have revealed that in such settings there is no hope to *recover* completely and accurately the predictive regions.

2.4 Current practices: thresholding decoding maps

Uniform threshold. Probably the most natural procedure used to recover discriminative patterns is to threshold decoders with high prediction performance —a popular choice is the linear SVM/SVR decoder [Pereira et al., 2009, Rizk-Jackson et al., 2011]. Thresholding decoder maps at a uniform value —*i.e.,* the threshold is the same for all weights— is probably the most common practice in neuroimaging; threshold value being generally arbitrary: "naked-eye criteria". It is not thought of as a statistical operation, and is sometimes left to the reader, who is presented unthresholded maps and yet told to interpret only the salient features of these maps.

Permutation testing can also be used to derive a uniform threshold with explicit guarantees. The classical Westfall-Young permutation test procedure [Westfall and Young, 1993] is well-known in the univariate context to control the FWER [Anderson, 2001], but its application to multivariate testing is not as straightforward. Then, instead of considering the usual t-statistics, a permutation test can use the linear SVR weights. An estimated weight map must be computed for the original problem and for several permuted problems before performing the Westfall-Young procedure; this method is detailed in Sec. 3.3.

Under some assumptions (see Sec. 3.2 and Sec. 3.3) that are more or less problematic in practice, the uniform thresholding strategies might recover the predictive patterns with FWER control. However, we will see that these naive strategies are not satisfactory in practice.

Non-uniform threshold. Another method proposed by Gaonkar and Davatzikos [2012], specifically designed for neuroimaging settings, relies on the analytic approximation of a permutation test performed over a linear SVM/SVR estimator. This method computes confidence intervals around the weights of the proposed estimator. Then, under some assumptions (see Sec. 3.4) that are not always met in practice, this procedure controls the FWER. It is almost equivalent to thresholding the SVR weights with a non-uniform threshold —*i.e.,* the threshold is specific to each weight. We refer to it as Adaptive Permutation Threshold SVR (Ada-SVR) from now on.

2.5 Building decoders designed for statistical control

Dimension reduction by voxel grouping. A computationally attractive solution to alleviate high dimensionality is to leverage the data structure and group adjacent —and

correlated— voxels, producing a closely related, yet compressed version of the original problem. In decoding, the grouping of voxels via spatially-constrained clustering algorithms has already been used to reduce the problem dimension [Gramfort et al., 2012, Varoquaux et al., 2012, Wang et al., 2015]. Specifically, groups of contiguous voxels can be replaced by the average signal they carry, reducing the dimensionality while improving the conditioning of the estimation problem. However, such a compression introduces a bias, as the patterns are constrained by the clusters shape. This bias is problematic as there is no unique grouping or clustering of the voxels [Thirion et al., 2014]: many different groupings capture the signal as accurately. One way to mitigate this bias is to use aggregation of models [Breiman, 1996, Zhou, 2012] obtained from several voxel groupings. Varoquaux et al. [2012] implemented this idea by computing different groupings from different random subsamples of the full data sample. The corresponding procedure yields decoders with more stable maps as well as a better prediction accuracy. In this subsampling spirit, random subspace methods [Ho, 1998, Kuncheva et al., 2010, Kuncheva and Rodríguez, 2010] also improve the prediction accuracy with more stable solutions —but in this case the subsampling is performed on the raw features. More recently, a procedure, *Fast Regularized Ensembles of Models* (FReM) [Hoyos-Idrobo et al., 2018], has combined clustering and ensembling to reduce the variance of the weight map, while ensuring high prediction accuracy. Yet, FReM weight maps do not enjoy statistical guarantees.

High-dimensional statistics tools. There have been a variety of procedures to produce p-value maps (map of p-values associated to every covariate) for linear models in high dimension [Wasserman and Roeder, 2009, Meinshausen et al., 2009, Bühlmann, 2013, Zhang and Zhang, 2014, Javanmard and Montanari, 2014]. Yet, they are not directly applicable to brain-imaging settings, as the dimensionality is too high. Based on a comparative review of those procedures [Dezeure et al., 2015], we have focused on the so-called Desparsified Lasso (DL), introduced in Zhang and Zhang [2014] and thoroughly analyzed by van de Geer et al. [2014]. Roughly, Desparsified Lasso can be seen as a Lasso-type [Tibshirani, 1996] extension of the least-squares to high dimensional settings, producing weight maps with well-controlled satistical distribution.

However, when the number p of features is much greater than the number n of samples, Desparsified Lasso lacks statistical power [Chevalier et al., 2018] and the computational cost becomes prohibitive. Indeed, solving Desparsified Lasso entails solving p Lasso problems with a design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$. Using the standard coordinate descent implementation [Friedman et al., 2007] the computation time is $\mathcal{O}(Tnp^2)$, with T the number of epochs used to solve the Lasso. However, when p is of order of few thousands and n few hundreds, Desparsified Lasso remains feasible with modest computer resources. In this context, the recently proposed Ensemble of Clustered Desparsified Lasso (EnCluDL) [Chevalier et al., 2018] combines three steps: a clustering procedure that reduces the problem dimension but preserves data structure, the Desparsified Lasso procedure that is tractable on the compressed problem, and an ensembling method introduced by Meinshausen et al. [2009] that aggregates several solutions of the compressed

problem. This method, summarized in Sec. 3.5, follows a scheme similar to FReM but the inference and ensembling procedures are different since it aims at producing p-value maps with statistical properties. Indeed, under some assumptions (see Sec. 3.5), it can be shown that EnCluDL controls the δ-FWER at the desired nominal level.

Finally, Knockoff filters [Barber and Candès, 2015, Candès et al., 2018], extended to work on images by Nguyen et al. [2019], are also an appealing procedure, though they can only control the FDR [Barber and Candès, 2015] or a relaxed version of the FWER [Janson and Su, 2016] incompatible with our spatial control, the δ-FWER detailed below. In this study, following the previous work of Chevalier et al. [2018], we focus on FWER or δ-FWER control. We then defer the extension of EnCluDL to FDR-controlling procedures and the benchmarking with alternatives to future work.

3 Materials and methods

3.1 δ-Family Wise Error Rate (δ-FWER)

In this section, we introduce a new way of controlling false detections that is well suited for neuroimaging settings as it incorporates spatial tolerance.

True support under linear model assumption. When considering multivariate inference, the *support* $S \subset [p]$ is the set of covariates that are non-independent of $\mathbf{y}$ conditionally to the other covariates. The rest of the voxels form the *null region* N *i.e.,* $N = [p] \setminus S$. Formally, S is the unique set that verifies:

$$\begin{aligned} &\forall j \in S, \qquad \mathbf{X}_j \not\perp\!\!\!\perp \mathbf{y} \mid \{\mathbf{X}_k, k \in [p] \setminus \{j\}\} \ , \\ &\forall j \in N, \qquad \mathbf{X}_j \perp\!\!\!\perp \mathbf{y} \mid \{\mathbf{X}_k, k \in S\} \ , \end{aligned} \tag{2}$$

where the sign $\perp\!\!\!\perp$ denotes independence. Under the linear assumption made in (1), S becomes simply the set of non zero weights and N the set of zero weights:

$$\begin{aligned} S &= \{j \in [p] : w_j^* \neq 0\} \ , \\ N &= \{j \in [p] : w_j^* = 0\} \ . \end{aligned} \tag{3}$$

δ-neighborhood. The variables $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p$ can also be characterized by the spatial proximity of their underlying voxels in brain space: given $\delta \geq 0$, a voxel $k \in [p]$ is in the δ-neighborhood of a voxel (or a set of voxels) if their distance is less than δ.

δ-null region. For $\delta \geq 0$, we denote by $S^{(\delta)}$ the δ-dilation of the support S, *i.e.,* the set of voxels in S or in its δ-neighborhood. By definition, $S \subset S^{(\delta)}$. We denote by $N^{(-\delta)}$ the δ-erosion (inverse operation of a δ-dilation) of the null region N, implying that $N^{(-\delta)} \subset N$. From the definition of N we have immediately:

$$N^{(-\delta)} = [p] \setminus S^{(\delta)} \ , \tag{4}$$

We refer to $N^{(-\delta)}$ as the δ-null region. As shown in Fig. 1, we interpret the δ-null region as the subset of the covariates which are at a distance less than δ from the support covariates. We also give a practical example of the δ-null region in the case of real fMRI data in appendix in Fig. 14.

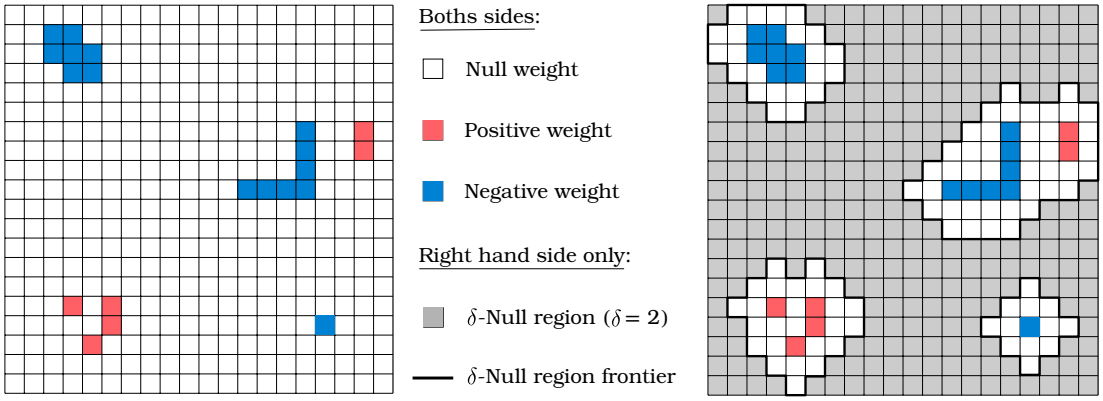


Figure 1: **Spatial tolerance to false discoveries.** Left: example of 2D-weight map, small squares represent voxels. The map is sparse. Right: representation of the δ-null region for the associated map with $\delta = 2$. The covariates in the δ-null region are "far" from non-null covariates, discoveries in this area are highly undesired. Discovering a null covariate "close" to a non-null covariate is tolerated.

δ-Family Wise Error Rate (δ-FWER). If we have an estimate of the support $\hat{S} \subset [p]$, we recall that the Family Wise Error Rate (FWER) is defined as the probability of making a false detection [Hochberg and Tamhane, 1987]:

$$\text{FWER}(\hat{S}) = \mathbb{P}(\hat{S} \cap N \neq \varnothing) \ . \tag{5}$$

Similarly, given $\delta \geq 0$, we defined the δ-FWER to be

$$\delta\text{-FWER}(\hat{S}) = \mathbb{P}(\hat{S} \cap N^{(-\delta)} \neq \varnothing) \ , \tag{6}$$

i.e., the probability of making a detection at distance more than δ from the true support. The δ-FWER control is thus weaker than the FWER control, except when $\delta = 0$ and when the true support is empty (*i.e.,* $N = [p]$), in which case the δ-FWER coincides with the classical FWER.

3.2 Thresholded SVR (Thr-SVR)

In this section, we introduce Thresholded SVR (Thr-SVR), a procedure that thresholds uniformly the estimated SVR weight map, keeping extreme weights; this method corresponds to the most standard and simple approach to recover predictive patterns. The first step is to derive the SVR weights $\hat{\mathbf{w}}^{\text{SVR}}$. Then, assuming that the estimated weights

of the null region are sampled from a given distribution centered on 0, the corresponding standard deviation σ_{SVR} can be approximated with the following estimator:

$$\hat{\sigma}_{\text{SVR}} = \sqrt{\frac{1}{p}\sum_{j=1}^{p}(\hat{\mathbf{w}}_j^{\text{SVR}})^2} \ . \tag{7}$$

We could also consider other estimators to approximate this quantity (*e.g.,* Schwartzman et al. [2009]) but the former is simple and at worst biased upward when the support is not empty. Now, assuming a Gaussian distribution for the SVR weights in the null region, *i.e.,* for $j \in N$:

$$\hat{\mathbf{w}}_j^{\text{SVR}} \sim \mathcal{N}\left(0, \sigma^2_{\text{SVR}}\right) \ , \tag{8}$$

we can produce (corrected) p-values by applying a Bonferroni correction. The produced p-values are at worst conservative under the two assumptions discussed in Section 6. In this procedure, the regression method considered is a linear SVR but similar results were obtained with other procedures (*e.g.,* Ridge regression).

3.3 Permutation Test SVR (Perm-SVR)

Now, we introduce another uniform thresholding strategy of SVR weights based upon a permutation test procedure. To derive corrected p-values from a permutation test, we first regress the design matrix against the response vector using a linear SVR to obtain an estimate $\hat{\mathbf{w}}^{\text{SVR}}$ of the weights map similarly as made in the Thr-SVR procedure. Then, permuting randomly R times the response vector and regressing the design matrix against the permuted response by a linear SVR, we obtain R maps $(\hat{\mathbf{w}}^{\text{SVR},(r)})_{r\in[R]}$. We can now apply the Westfall-Young step-down maxT adjusted p-values algorithm [Westfall and Young, 1993, p. 116-117] taking the raw SVR weights instead of the usual t-statistics to derive the corrected p-values. A sufficient assumption to ensure the validity of the p-values is the pivotality of the SVR weights. Keeping the corrected p-values that are less than a given significance level —equal to 10% in this study— this procedure is equivalent to thresholding the SVR weight map. We call this procedure Permutation Test SVR (Perm-SVR). The only difference between Perm-SVR and the Thr-SVR procedure is the way of computing the threshold. To perform the permutation test procedure, we took $R = 1000$ permutations.

3.4 Adaptive Permutation Threshold SVR (Ada-SVR)

Here, we introduce Adaptive Permutation Threshold SVR (Ada-SVR), a statistical inference procedure that produces a weight map and confidence intervals around it; it is also almost equivalent to thresholding the SVR weights non-uniformly. Ada-SVR was first presented by Gaonkar and Davatzikos [2012]. First, the authors derived an estimated weight $\hat{\mathbf{w}}^{\text{APT}}$ linearly related to the target by approximating the hard margin SVM formulation, their estimator is given by the following equation:

$$\hat{\mathbf{w}}^{\text{APT}} = \mathbf{L}\,\mathbf{y} \ , \tag{9}$$

where $\mathbf{y}$ is the target variable and $\mathbf{L} \in \mathbb{R}^{p \times n}$ only depends on the design matrix $\mathbf{X}$:

$$\mathbf{L} = \mathbf{X}^\top \left[(\mathbf{X}\mathbf{X}^\top)^{-1} - (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{1}(\mathbf{1}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{1})^{-1}\mathbf{1}^\top(\mathbf{X}\mathbf{X}^\top)^{-1} \right] \ , \tag{10}$$

where $\mathbf{1} \in \mathbb{R}^n$ is a vector of ones. The approximation made by (9) is notably valid under the assumption that all the data samples are support vectors, which might hold at least if $n \ll p$. Then, if $\mathbf{y}$ is standardized and if n is large enough (so that the central limit theorem holds), one expects that under the null hypothesis for the j-th covariate:

$$\hat{\mathbf{w}}_j^{\mathrm{APT}} \sim \mathcal{N}(0, \sum_{k=1}^{n} \mathbf{L}_{j,k}^2) \ . \tag{11}$$

From (11), p-values can be computed and corrected by applying a Bonferroni correction (multiplying the raw p-values by a factor p).

3.5 Ensemble of Clustered Desparsified Lasso Algorithm (EnCluDL)

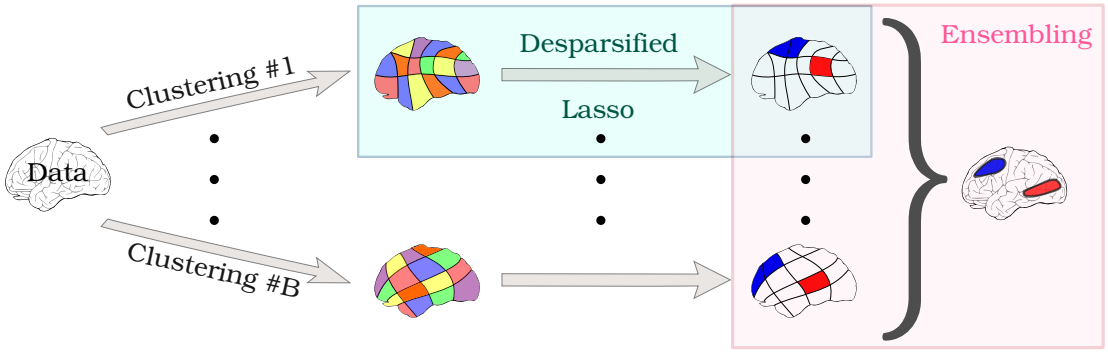


Figure 2: **Ensemble of Clustered Desparsified Lasso (EnCluDL) algorithm.** The EnCluDL algorithm combines three algorithmic steps: a clustering (or parcellation) procedure applied to images, the Desparsified Lasso procedure (statistical inference) to derive statistical maps, and an ensembling method that synthesizes several statistical maps. In the first step, B clusterings of voxels are generated using B random subsamples of the original sample. Then, for each grouping-based data reduction, a statistical inference procedure is run resulting in B z-score maps (or p-value maps). Finally, these maps are ensembled into a final z-score map using an aggregation method that preserves statistical properties.

Ensemble of Clustered Desparsified Lasso (EnCluDL) is a multivariate statistical inference procedure designed for spatial data; it was first introduced by Chevalier et al. [2018]. EnCluDL relies on three steps: a spatially-constrained clustering algorithm for reducing the problem dimension, a statistical inference procedure for deriving statistical maps, and an ensembling method for aggregating the statistical maps.

Statistical inference with Desparsified Lasso. Desparsified Lasso (DL) is a statistical inference procedure that can be viewed as a generalization of the least-squares-based inference in high dimension under sparsity assumptions. It was proposed and thoroughly analyzed by Zhang and Zhang [2014] and van de Geer et al. [2014]. This estimator produces p-values on linear model parameters even when the number of parameters p is (reasonably) greater than the number of samples n. A technical description of Desparsified Lasso is available in Sec. 7.1. In the neuroimaging context, the initial parameters are related to the voxels, which are of the order of one hundred thousand while the number of samples is almost always lower than one thousand. In such settings Desparsified Lasso is inefficient due to a lack of statistical power, hence dimension reduction is required.

Clustering. As argued in Section 1, while performing dimension reduction, we aim at keeping the spatial structure of the data and avoid mixing voxels "far" from each other. This is achieved with data-driven parcellation along with a spatially constrained clustering algorithm following the conclusions by Varoquaux et al. [2012] and Thirion et al. [2014]. Another interesting aspect of this dimension reduction method is its denoising property [Hoyos-Idrobo et al., 2018] since it produces averages from groups of noisy voxels. Note that this choice ultimately calls for a spatial tolerance on the statistical control, *i.e.,* considering the δ-FWER instead of the standard FWER. Through the clustering, the p voxels are grouped into C clusters, where $C \ll p$. Then, Desparsified Lasso is directly applied to the compressed problem in order to produce corrected p-values. Notably, corrected p-values are obtained from the initial p-values by applying Bonferroni correction [Dunn, 1961] with a factor $C \ll p$. Following the terminology in [Chevalier et al., 2018], we refer to this procedure as Clustered Desparsified Lasso (CluDL). CluDL however suffers from high variance [Chevalier et al., 2018] as it depends on an arbitrary grouping choice. This can be alleviated by ensembling techniques, as described next.

Ensembling. Varoquaux et al. [2012], Hoyos-Idrobo et al. [2018] have shown that randomizing the grouping choice and adding an ensembling step to aggregate several solutions can stabilize the overall procedure. Additionally, Chevalier et al. [2018] have highlighted that the ensembling step is also beneficial in terms of support recovery. To perform B groupings of the covariates, we train the parcellations algorithm with B different random subsamples of the original data sample. Then, thanks to the CluDL procedure, we obtain B statistical maps that are aggregated into one through an ensembling procedure. The ensembling procedure we considered in the statistical inference procedure is adapted from Meinshausen et al. [2009] that is described in appendix in Sec. 7.2. We refer to the full inference algorithm as Ensemble of Clustered Desparsified Lasso (EnCluDL). Under hypothesis ensuring Desparsified Lasso statistical properties —notably sparsity and smoothness of the true weight map and i.i.d. data samples— EnCluDL gives statistical guarantees, namely it controls the δ-FWER.

Choosing δ for δ-FWER control Theoretically, the minimal spatial tolerance δ that guarantees a control of the δ-FWER with EnCluDL is the largest parcel diameter.

However, in practice, we aggregate many statistical maps obtained from different choices of voxel grouping; then the required spatial tolerance is reduced to the average radius. Then, the value of δ for which we observe the δ-FWER control varies approximately linearly with the cubic root of the average number of voxels per cluster. In standard fMRI settings, we propose the following formula for δ:

$$\delta_0 = \left(\frac{p}{2C}\right)^{1/3} , \tag{12}$$

the ratio p/C being the average number of voxels per cluster, δ_0 is a distance in voxel size unit.

Note that the previous formula is an estimate of the average cluster radius that assumes that the shape of the clusters have identical cubic shape. In practice, this formula tends to underestimate the average cluster radius but was suitable in all our experiments. In Sec. 7.6, we study empirically the distribution of the cluster radius distribution as a function of the number of clusters, and compare it with δ_0.

Additionally, note that when the setting is particularly favorable for inference, *e.g.,* if $\log(n)/C$ is large, the choice of δ given by (12) might be slightly too liberal. To address these specific cases, we propose a more refined formula to estimate δ in appendix in Sec. 7.5.

EnCluDL hyper parameters. The number of clusters C is a crucial hyperparameter of EnCluDL. Generally, a suitable C depends on intrinsic physical properties of the problem and on the targeted spatial tolerance δ. Decreasing C increases the statistical power while reducing the spatial precision. In the neuroimaging context, taking $C = 500$ is a fair default value achieving a suitable trade-off between spatial precision and statistical power when the number of samples is a few hundreds. With this choice, the spatial tolerance should be close to $\delta = 10\,\text{mm}$ when working with masked fMRI data.

As a more adaptive approach, we recommend tuning C according to n *e.g.,* $C \in [n/2, n]$. This choice should still ensure the δ-FWER control with δ given by (12) (or its corrected version, see appendix Sec. 7.5) and is justified in Sec. 4.5.

The parameter B, the number of CluDL solutions to be aggregated, is discussed in Sec. 3.5. The larger B the more stable the solution, yet the heavier the computational cost. In our experiments, we have set $B = 25$ (see Hoyos-Idrobo et al. [2018] for a more complete discussion on this parameter).

Empirical analysis of data structure assumptions for EnCluDL. The core part of EnCluDL consists in applying Desparsified Lasso to a clustered version of the original problem. As disclaimed in van de Geer et al. [2014], some technical hypotheses on the structure of the design matrix $\mathbf{X}$ —*i.e.,* of the reduced data— are necessary to produce valid confidence intervals on the parameters with Desparsified Lasso. Roughly, it is necessary that the features are "not too much correlated". In appendix in Sec. 7.3, we show in a simple setting that as long as the correlation between two predictive features

is less than 0.8, it is possible to recover both features. However when the correlation between features is more than 0.9, only one of the two features can be identified.

In Sec. 7.4, we show that in standard fMRI datasets neighboring voxels can have a correlation greater than 0.9. Thus applying Desparsified Lasso at the voxel level certainly leads to many false negatives. However, since Desparsified Lasso is applied to the clustered problem, we have to consider correlation between clusters instead. In Sec. 7.4, we show on HCP data that such inter-cluster correlation is almost always lower than 0.8 and always lower than 0.85. This means that data structure assumptions for EnCluDL are sustainable. Additionally, the fact that EnCluDL aggregates several CluDL solutions increases the tolerance to inter-cluster correlation.

3.6 A complementary univariate solution

Given the complementarity of univariate and multivariate inference noted previously, we add to our study a univariate inference method, namely *univariate permuted OLS* (Univ-OLS). This method does not test the same null hypothesis as the other methods: it tests whether or not a voxel is marginally associated with the target. Then, while it should not be benchmarked with the other methods, we propose to consider jointly the results obtained by the marginal and the conditional analyses, as advocated by Weichwald et al. [2015].

The Univ-OLS method is based on the generalized linear model (GLM) [Friston et al., 1994]. For every voxel we compute a t-statistic by applying the OLS procedure on the linear model that associates each voxel with the target. Subsequently, we also derive the permuted t-statistic distribution by performing the OLS on permuted data. Finally, to obtain corrected p-values, we use the standard maxT procedure [Westfall and Young, 1993]. Note that, for this method, we have used the `permuted_ols` function implemented in the Nilearn python package [Abraham et al., 2014] with 1000 permutations.

3.7 Implementation

The Python code that implements Thr-SVR, Perm-SVR, Ada-SVR and EnCluDL can be found on `https://github.com/ja-che/hidimstat`. Our algorithms are implemented with Python = 3.6.8 and need the following packages Numpy = 1.16.2 [Van der Walt et al., 2011], Scipy = 1.2.1 [Virtanen et al., 2020], Scikit-Learn = 0.21 [Pedregosa et al., 2011], Joblib = 0.11 and Nilearn = 0.6.0 [Abraham et al., 2014].

4 Experimental procedures

4.1 Data

To validate empirically the statistical guarantees of the four algorithms —Thr-SVR, Perm-SVR, Ada-SVR and EnCluDL— described in Section 3, we perform several experiments on resting-state fMRI and task fMRI data. We also show some results for Univ-OLS to highlight the complementarity of univariate and multivariate analyses, in

particular when studying predictive patterns on real data. We focus on three datasets: HCP900 resting-state fMRI, HCP900 task fMRI and RSVP task fMRI.

HCP900 resting-state fMRI data. HCP900 resting-state fMRI dataset [Van Essen et al., 2012] contains 4 runs of 15 minutes resting-state recordings with a 0.76s-repetition time (corresponding to 1200 frames per run) for 796 subjects. We use the MNI-resampled images provided in the HCP900 release. For this dataset the number of samples is equal to 1200 (only one run is used) and the number of voxels is 156 374 after gray-matter masking (the spatial resolution being 2 mm isotropic).

HCP900 task fMRI data. We also use the HCP900 task-evoked fMRI dataset [Van Essen et al., 2012], in which we take the masked 2 mm z-maps of the 796 subjects from 6 tasks to solve 7 binary classification problems: emotion (*emotional face* vs *shape outline*), gambling (*reward* vs *loss*), language (*story* vs *math*), motor hand (*left* vs *right* hand), motor foot (*left* vs *right* foot), relational (*relational* vs *match*) and social (*mental interaction* vs *random interaction*). We consider the fixed-effect maps for each outcome (or condition), yielding one image per subject per condition (which corresponds to two images per subject for each classification problem). Then, for each problem, the number of samples available is 1592 ($= 2 \times 796$) and the number of voxels is 156 374 after gray-matter masking.

Unmasked RSVP task fMRI data. We also use activation maps obtained from a rapid serial visual presentation (RSVP) task of the individual brain charting dataset [Pinho et al., 2018], augmented with 9 additional subjects performing the same task, under the same experimental procedures and scanning parameters. No masking is used for this dataset, so that out-of-brain voxels are not withdrawn from preprocessing. We consider the unmasked 3 mm-resolution statistical z-maps of the 6 sessions of the 21 subjects for a reading task with 6 different contrasts that have been grouped into 2 classes: language (words, simple sentences, complex sentences) vs pseudo-language (consonant strings, pseudo-word lists, jabberwocky). The images are all registered to MNI space and per-condition effects are estimated with Nistats v0.0.1 library [Abraham et al., 2014]. For this dataset the number of samples available is equal to 756 (21 subjects $\times$ 6 runs $\times$ 6 images per run) and the number of voxels is 173 628 (unmasked images resampled at 3-mm resolution). We run the inter-subject experiment described in Sec. 4.4 with this dataset.

4.2 Statistical control on semi-simulated data

A first series of experiments study whether the four different methods exhibit the expected δ-FWER control and are competitive in terms of support recovery, as measured with the precision-recall curve. To do so, we have to construct the true weight map $\mathbf{w}^*$. We generate "semi-simulated" data: generating signals from estimates on real data. To

avoid circularity in the definition of the ground truth, we used two different tasks: one to build $\mathbf{w}^*$ and another one to define $\mathbf{X}$.

Building a reference weight map from HCP900 motor hand dataset. To construct an underlying weight map, we use the motor hand (MH) task of the HCP900 task fMRI dataset described in Sec. 4.1. Specifically, we build a design matrix $\mathbf{X}_{\text{MH}} \in \mathbb{R}^{n \times p}$ from the motor hand task z-maps of all subjects associated with a binary target index $\mathbf{y}_{\text{MH}}$. To obtain an initial weight map $\mathbf{w}_{\text{MH}}^{\text{SVC}}$ we regress $\mathbf{X}_{\text{MH}}$ against $\mathbf{y}_{\text{MH}}$ by fitting a linear Support Vector Classifier (SVC) [Cortes and Vapnik, 1995]. From $\mathbf{w}_{\text{MH}}^{\text{SVC}}$ we only kept the 10% most extreme values ensuring that the connected groups of non zero-weight voxels have a minimal size of 1 cm^3 by removing small clusters. We chose this map (represented in Fig. 3 and Fig. 4) to be the true weight map $\mathbf{w}^* \in \mathbb{R}^p$ for the whole simulated experiments.

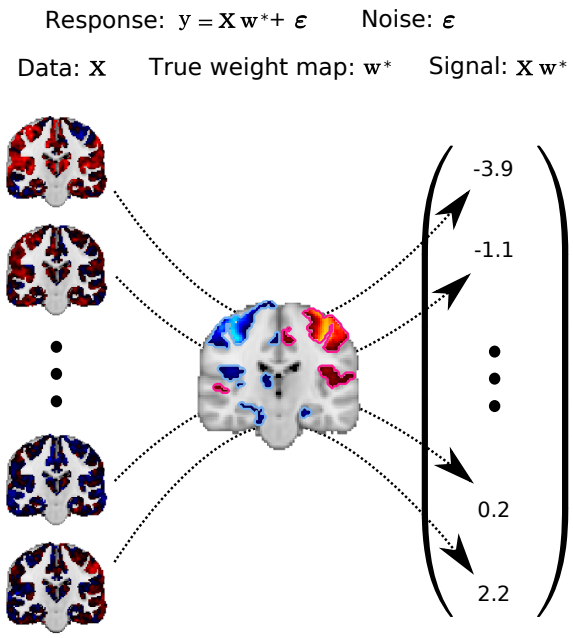


Figure 3: **Generating a hybrid dataset with known ground truth and actual fMRI data.** To generate the response for a given sample we multiply the corresponding brain activation map by the true weight map and add a Gaussian noise with fixed variance. To highlight the predictive regions, we circle them in pink for positive coefficients and in light blue for negative coefficients. As an illustration, we take four different data samples with negative or positive output value.

Simulating responses with HCP900 emotion dataset. We then take $\mathbf{X}$ to be the set of z-maps from the emotion task of the HCP900 task fMRI dataset described in Sec. 4.1. To generate a continuous response vector $\mathbf{y}$, we draw a Gaussian random noise vector $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_n)$ and use the linear model introduced in (1), where $\sigma_\varepsilon = 0.2$ to reach $\text{SNR}_y = 10$, where SNR_y is given by:

$$\text{SNR}_y = \frac{\|\mathbf{X}\,\mathbf{w}^*\|^2}{n\sigma_\varepsilon^2} \quad . \tag{13}$$

The way we simulate $\mathbf{y}$ is summarized in Fig. 3.

Quantification of error control and detection accuracy. To obtain representative results, we then run the procedures described in Section 3 for 100 different response vectors $\mathbf{y}$ generated from different random samples of subjects and different draws of ε. We let the number of samples vary from $n = 50$ (25 random subjects taken among the 796) to $n = 1200$ (600 subjects), the number of voxels being $p = 156\,374$. For each simulation, we record the empirical δ-FWER and the precision-recall curves. Importantly, we do not recommend running such analysis with $n < 100$, since the estimation problem is hard and statistical guarantees are only asymptotic.

Heavy-tailed version of the semi-simulated experiment. In the above experiment the noise is Gaussian, hence we also benchmark the inference procedures for Laplace and Student noise to assess the impact of noise distribution.

Binary version of the semi-simulated experiment. In the main experiment the response vector $\mathbf{y}$ is continuous, hence we also benchmark the inference procedures for a binary response. For that, we simply take as response vector the signs of the continuous $\mathbf{y}$ generated as in the previous paragraph.

Univ-OLS solves another inference problem. Univariate methods do not compete with multivariate methods, as they do not test the same null hypotheses. However, for pedagogical purpose, we show that Univ-OLS based FWER control is not valid in the multivariate analysis setup.

4.3 Statistical control under the global null with i.i.d. data

In this experiment, we test whether the procedures control the FWER under a global null model. EnCluDL only controls the δ-FWER theoretically but, when the true weight vector $\mathbf{w}^*$ is null, the δ-FWER and the classical FWER are identical. Then, all procedures should control the FWER. Here, we considered the tasks of the HCP900 task fMRI dataset described in Sec. 4.1 keeping all the subjects ($n = 1592$). Then, to get a noise-only response, we (uniformly) randomly permute the original response vector. Similarly as in Sec. 4.2, the i.i.d. hypothesis is legitimate, since the data correspond to z-maps of different subjects. For each task, we draw 100 different permutations of the response and check if the different methods enforce the chosen nominal FWER of 10%.

to illustrate the importance of checking the underlying assumptions, in appendix in Sec. 7.8, we describe an additional experiment to show that FWER (or δ-FWER) is not controlled anymore when working with an autocorrelated response vector, breaking the i.i.d hypothesis. This experiment is adapted from Eklund et al. [2016].

4.4 Statistical control of out-of-brain detections

In this experiment we test the four procedures on an unmasked task fMRI dataset to verify that no spurious detection is made outside of the brain —up to the allowed error rate. Indeed, the non-null coefficients of the weight vector $\mathbf{w}^*$ should all be contained

in the brain since there is no informative signal in out-of-brain voxels. To do so, we take the unmasked RSVP task fMRI dataset, described in Sec. 4.1 (with design matrix $\mathbf{X}$ containing $n = 756$ unmasked z-maps). Then, we report how frequently some voxels are detected outside the brain volume. For the sake of completeness, we also check the non-occurrence of out-of-brain detections with Univ-OLS.

4.5 Insights on the choice of number of clusters

In this experiment, we assess empirically the impact of C, the number of clusters used in the EnCluDL algorithm. We use the same generative method as in Sec. 4.2 to produce an experiment with known ground truth. Then, we run the EnCluDL algorithm varying the numbers of clusters C from $C = 200$ to $C = 1000$. We also vary the number of samples n from 100 to 1200. As in Sec. 4.2, we run the experiment for 100 different response vectors and report aggregated results. We report two statistics: the empirical δ-FWER and the AUC of the precision-recall curve for every value of C and n.

4.6 Face validity on HCP dataset

In this experiment, we consider the output of the procedures in terms of brain regions that are conditionally associated with the task performed by the subjects. Similarly as in Sec. 4.3, we consider the tasks of the HCP900 task fMRI dataset described in Sec. 4.1, keeping this time the true response vector. We run all the procedures on every task and report the statistical maps thresholded such that the FWER $< 10\%$ or the δ-FWER $< 10\%$ (for EnCluDL). For this, we use all the available samples ($n = 1592$). We also include Univ-OLS to compare the discriminative patterns obtained with a univariate inference.

4.7 Prediction performance

Even if it is not the purpose of this study, we also checked the prediction performance of the decoders produced by each method. Since Thr-SVR and Perm-SVR rely on the same predictive function, there are three different decoders: SVR, Ada-SVR and EnCluDL. To perform this experiment, we consider the tasks of the HCP900 task fMRI dataset described in Sec. 4.1. We run all the procedures on every task using a sample size $n = 400$, keeping the rest of the samples to test the trained model. For each task and each method, we take 100 different random subsamples to produce the results. This experiment being a side study, we give the results in appendix in Sec. 7.12.

5 Results

In this section, after setting the value of the tolerance parameter δ in the different datasets, we present the experimental results.

5.1 Estimating δ in HCP and RSVP datasets

In all the experiments, unless specified otherwise, we run EnCluDL with the default choice $C = 500$. Reversing (12), we obtain a tolerance parameter of $\delta_{\text{HCP}} = 5.4$ voxels for HCP900 and $\delta_{\text{RSVP}} = 5.6$ voxels for RSVP, corresponding to $\delta_{\text{HCP}} = 12\,\text{mm}$ and $\delta_{\text{RSVP}} = 18\,\text{mm}$ respectively after rounding up. In Fig. 14 in appendix, we display the spatial tolerance of 6 voxels in the case of HCP data.

5.2 Statistical control with known ground truth

Here, we describe the results obtained from the experiment described in Sec. 4.2.

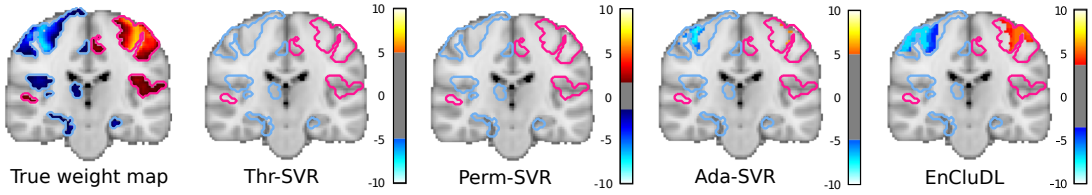


Figure 4: **Qualitative comparison of the model solutions.** Here, we show the solutions (z-maps) given by the four inference procedures, for a single random draw of the noise vector in the experiment described in Sec. 4.2. The weight maps are thresholded such that δ-FWER $< 10\%$ theoretically. We can observe that none of the methods yield false discoveries but the Ensemble of Clustered Desparsified Lasso (EnCluDL) procedure is the most powerful followed by Adaptive Permutation Threshold SVR (Ada-SVR).

Qualitative comparison of the model solutions. In Fig. 4, we present a qualitative comparison of the model solutions when $n = 400$. None of the methods yields false discoveries for the chosen threshold —taken such that δ-FWER $< 10\%$. EnCluDL recovers more active regions than the other procedures, which makes it the most powerful procedure, followed by Ada-SVR. The other two procedures do not discover the expected patterns. These results displayed are obtained for a single random draw of the noise vector, but similar results holds for different draws.

δ-FWER control. In this experiment, we check if Thr-SVR, Perm-SVR, Ada-SVR and EnCluDL control the δ-FWER at the targeted nominal level (here being 10%). Fig. 5 shows that Perm-SVR and EnCluDL procedures control the δ-FWER for all sample sizes since their empirical δ-FWER remain below the targeted nominal level, whereas Thr-SVR and Ada-SVR fail to control the δ-FWER in every setting. In particular, the empirical δ-FWER for Ada-SVR is above the targeted nominal level for $n \geq 800$. This might occur since the approximation made by (9) is valid only if n remains "sufficiently low" [Gaonkar and Davatzikos, 2012]. Thr-SVR fails to control empirically the δ-FWER for any value of n. This might be due to the two assumptions made in Sec. 3.2 not being satisfied —it is indeed unlikely that the SVR weights of the null region follow the same

distribution. We further discuss this point in Section 6. Concerning EnCluDL, one can notice that the empirical δ-FWER is slightly larger for $n = 1200$, this effect is explained in appendix in Sec. 7.5 and Sec. 7.6. We report additional results, notably heavy-tailed and binary version of the experiment, in appendix in Sec. 7.10. These lead to the same statistical behavior as observed here.

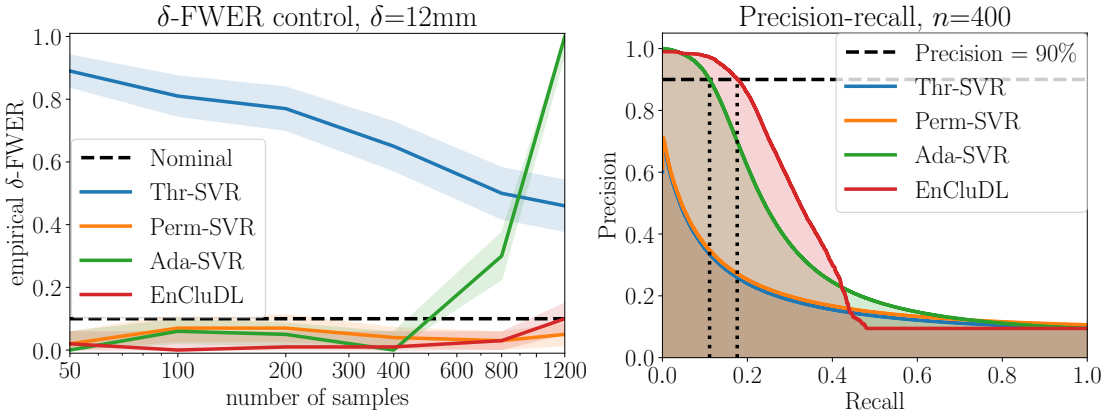


Figure 5: **δ-FWER control and precision-recall curve on semi-simulated data (known ground truth).** Left: The results of the experiment described in Sec. 4.2 show that the permutation test (Perm-SVR) and Ensemble of Clustered Desparsified Lasso (EnCluDL) are the only procedures that correctly control the δ-FWER at the nominal level (10%). This is not the case for Adaptive Permutation Threshold SVR (Ada-SVR) and Thresholded SVR (Thr-SVR) procedures. Right: For the same experiment, EnCluDL has the best performance in terms of precision-recall curve. For $n = 400$, and ensuring 90% precision, EnCluDL obtains a recall of 23% and Ada-SVR a recall of 16%. Thr-SVR and Perm-SVR share the same precision-recall curve and were not able to reach 90% precision.

Precision-recall. In this experiment, we also evaluate the recovery properties of the four methods by comparing the precision-recall curve for different value of n. Fig. 5 shows that EnCluDL has the best precision-recall curve for $n = 400$. We recall that the perfect precision-recall curve is reached if the precision is equal to 1 for any value of recall between 0 and 1. Similar results were obtained for the other sample sizes tested (appendix Fig. 17). Indeed, when $n = 400$, for a 90% precision, EnCluDL gives a recall of 23% and Ada-SVR a recall of 16%. Thr-SVR and Perm-SVR share the same precision-recall curve since they both produce p-values arranged in the reverse order of the absolute SVR weights. These thresholding methods were not able to reach the 90% precision; their recovery properties are much weaker.

We report additional results in Sec. 7.10.

5.3 Statistical control under the global null with i.i.d. data

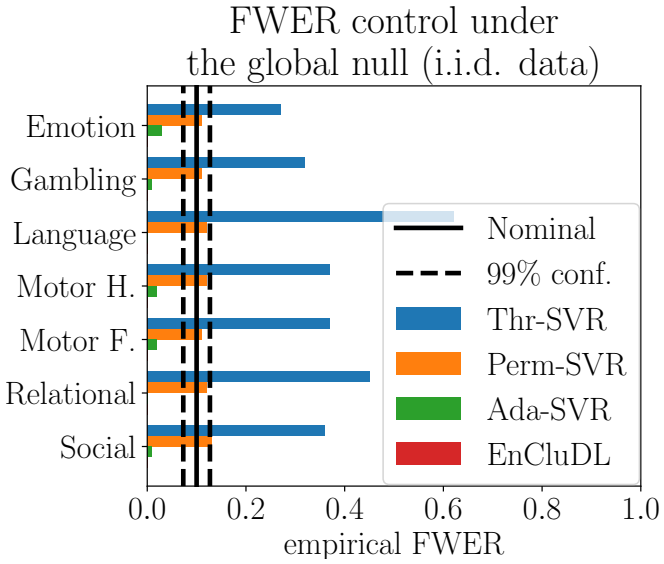
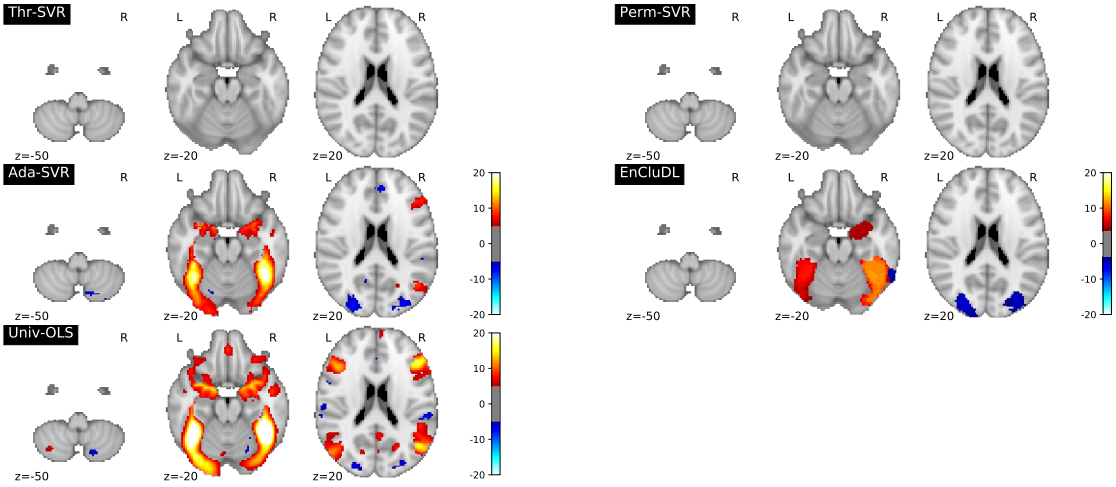


Figure 6: **FWER control under the global null with i.i.d. data** The results of the experiment with i.i.d. data under the global null, described in Sec. 4.3, show that, only the Thresholded SVR (Thr-SVR) fails to control the FWER empirically in this context. EnCluDL makes no detection: it is a conservative approach, as one could expect from theory.

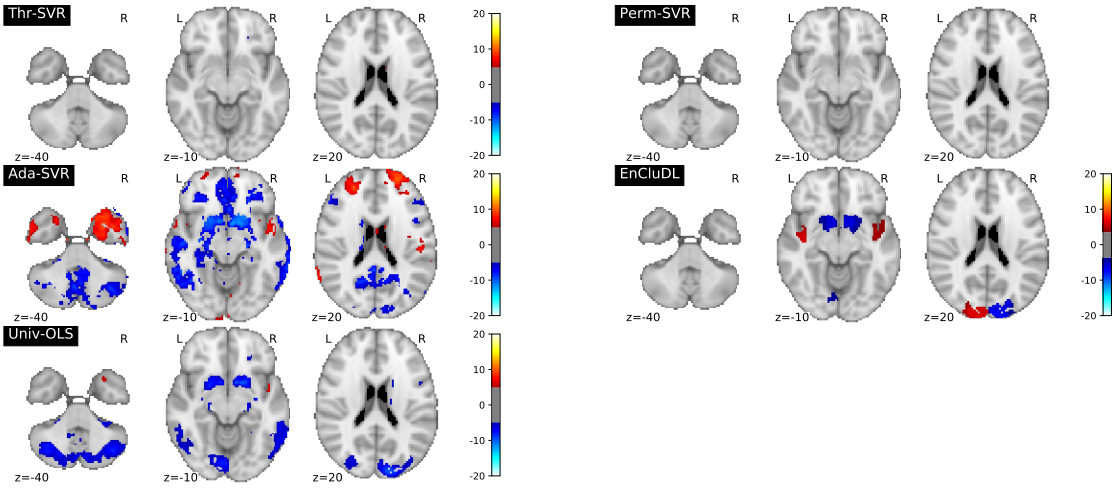
FWER control under the global null (permuted response). Here, we summarize the results of the experiment testing control of the FWER in a global null setting (Sec. 4.3). Fig. 6 shows that, when samples are i.i.d., all the procedures control the FWER, except Thr-SVR. EnCluDL is even conservative since the empirical FWER remains at 0 for all the different tasks tested. This result is not surprising since at least two steps of the EnCluDL procedure are conservative: the Bonferroni correction and the ensembling of the p-values maps.

Face validity (original response). Additionally, we run the procedures with the original (not permuted) response vector to check whether the methods can recover predictive patterns; this corresponds to the experiment described Sec. 4.6. We plot the results for the two first tasks (emotion and gambling) in Fig. 7; see appendix Fig. 23 for the five other tasks. Qualitatively, EnCluDL recovers the most plausible predictive patterns, Ada-SVR sometimes makes dubious discoveries: patterns are too wide and implausible. The two other methods exhibit a very weak statistical power.

Comparing EnCluDL and Univ-OLS solutions, we see that the discovered patterns are not a subset of each other. This result was expected given the arguments in Weichwald et al. [2015]: the advantage of combining the two paradigms is to get more insight on the causal nature of the relation between the voxel signals and the target.



(a) Emotion



(b) Gambling

Figure 7: **Estimated predictive patterns on standard task fMRI dataset.** Here, we plot the results for the emotion and gambling tasks of the experiment described in Sec. 4.6 thresholding the statistical maps such that the δ-FWER stays lower than 10% for $\delta = 12\,\text{mm}$. Qualitatively, EnCluDL discovers the most plausible patterns, Ada-SVR sometimes makes dubious discoveries, patterns are too wide and implausible, while the two other methods exhibit a very weak statistical power. Univariate analysis results obtain with Univ-OLS clearly provide distinct information about the relationship between the voxel signals and the outcome. The results of the five other tasks are available in Fig. 23.

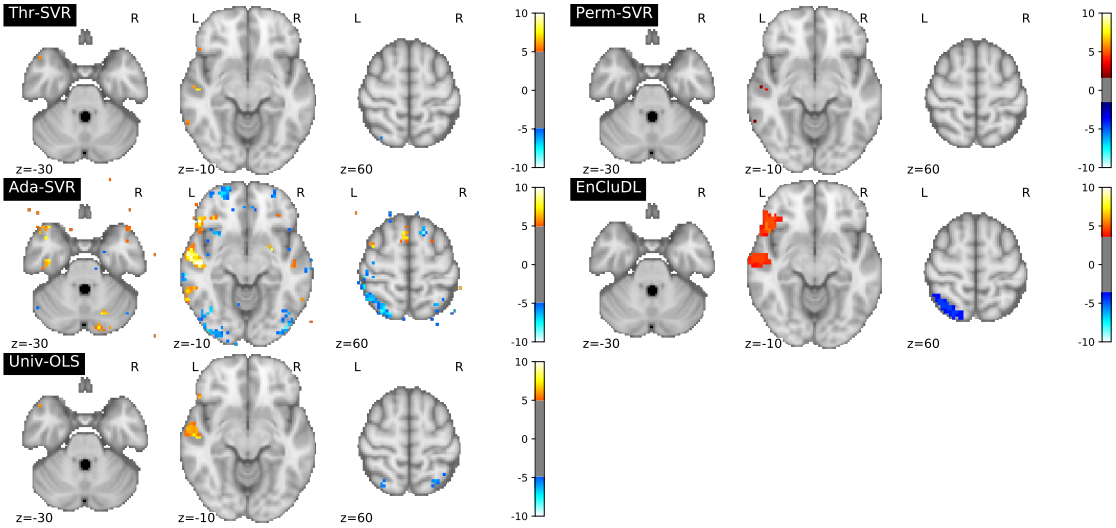


Figure 8: **Statistical maps for unmasked RVSP data.** The results of the unmasked task-fMRI experiment, described in Sec. 4.4, show that EnCluDL, Thresholded SVR (Thr-SVR) and the permutation test (Perm-SVR) do not return out-of-brain discoveries, while the Adaptive Permutation Threshold SVR (Ada-SVR) does. Here z-score maps are thresholded such that the δ-FWER is at most 10% for $\delta = 6$ voxels (or $18\,\text{mm}$). Thr-SVR and the Perm-SVR do not yield spurious detections but very few detections are made, hence these method have low statistical power. EnCluDL does not make any spurious detection; rather it makes detections in the temporal lobe and Broca's area, which are expected for a reading task. Univ-OLS does not make any out-of-brain detection either but returns significant associations in the temporal lobe.

5.4 Statistical control of out-of-brain discoveries

We now report the results from the unmasked RSVP task data experiment (Sec. 4.4). Here, we check whether out-of-brain detections are made. In Fig. 8, the z-score maps are thresholded such that the FWER (for Perm-SVR, Thr-SVR, and Ada-SVR) or the δ-FWER (for EnCluDL) are at most 10% for $\delta = 6$ voxels (or $18\,\text{mm}$). We observe that Ada-SVR makes some out-of-brain discoveries, and it does not control the FWER empirically. Thr-SVR and Perm-SVR do not yield spurious detections but very few detections are made, hence these methods have low statistical power. EnCluDL does not make any out-of-brain detections and it outlines predictive regions in the temporal lobe and Broca's area, expected for a reading task. Finally, Univ-OLS does not make any spurious detection either; it only makes detections in the temporal lobe.

5.5 Insights on choosing the number of clusters

Here, we report the results obtained of the experiment task-fMRI data (Sec. 4.5) studying the impact of C (number of clusters) on the δ-FWER control and the recovery

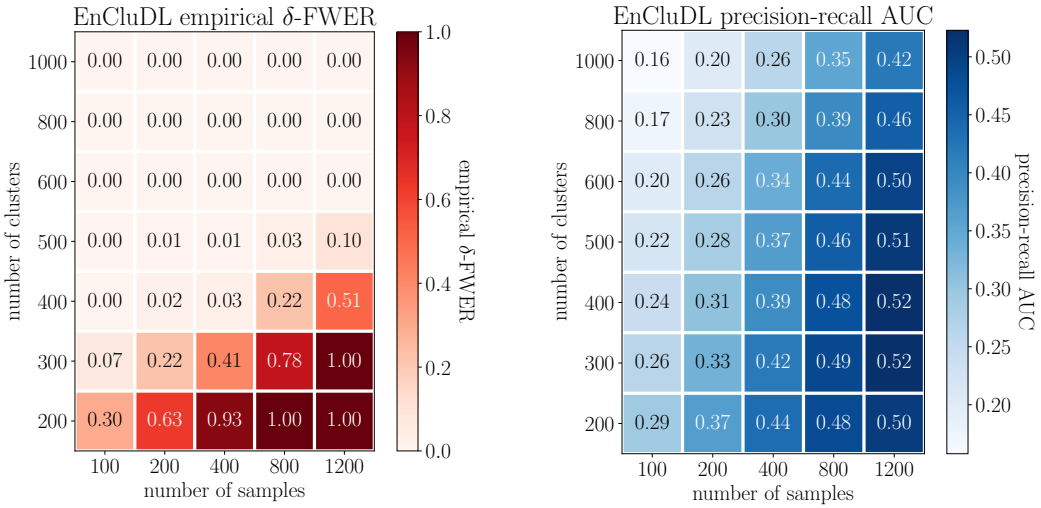


Figure 9: **Influence of the number C of clusters on δ-FWER control and the recovery properties of EnCluDL.** The results of the experiment described in Sec. 4.5 show the impact of C on the δ-FWER control and the recovery score of EnCluDL. When $C \geq 500$, clusters are smaller, hence the δ-FWER is controlled for $\delta = 12\,\text{mm}$ (and potentially lower values of δ) since all the empirical δ-FWER's are lower than the 10% nominal rate. Conversely, when $C < 500$, clusters are wider and the spatial tolerance is overcome by the model inaccuracy, hence the δ-FWER is not controlled for $\delta = 12\,\text{mm}$. However, it remains controlled for higher values of δ. Concerning the recovery properties we see that reducing the number of clusters improves the precision-recall curves. Thus, the more spatial uncertainty is tolerated, the best recovery properties EnCluDL offers.

properties of EnCluDL for various sample sizes. These results are obtained with 100 repetitions for every sample and cluster sizes. In Fig. 9, we notice that a lower C leads to improved recovery, according to the area under the precision-recall curves, for $\delta = 6$ voxels (or 12 mm). However, when the number of cluster is lower, the average cluster radius increases and overcomes the spatial tolerance of δ, leading to inflated error rates (cf. Sec. 7.6). More precisely, the δ-FWER is controlled when $C \geq 500$. Note that for $C < 500$, it is possible to control the δ-FWER, even when n is small, provided a larger spatial tolerance $\delta > 6$ voxels. To compute the requested δ, one can use (12). Besides, we observe that the recovery score of EnCluDL improves when n increases, as expected. We also notice that the empirical δ-FWER increases with n. To explain this effect, we first recall that theoretically the δ-FWER is controlled for δ equal to the largest cluster diameter, likely to be too large in practice. In this study, we have taken δ equal to δ_0, which is slightly smaller than the average radius of the clusters (cf. Sec. 7.6), since in practice this choice ensures the δ-FWER control. However, when the setting is particularly favorable for inference (*e.g.*, if $\log(n)/C > 1.5 \times 10^{-2}$), some false discoveries can be made at a distance greater than the average radius from the support. The choice of δ

is further discussed in Sec. 3.5 and in appendix in Sec. 7.5. Additionally, we can notice from Fig. 9 that for a fixed C/n ratio the recovery capability is stable (see also appendix Sec. 7.9). Then, as discussed in Sec. 3.5, we advise taking C of the same order as n (*e.g.*, $C \in [n/2, n]$) when the goal is to recover most of the predictive regions without strict requirements on the accuracy of their shapes —since the value of δ given by (12) might be not small with regards to the predictive region itself.

6 Discussion

Decoding models are fundamental for causal interpretation of the implication of brain regions for an outcome of interest, mental process or disease status [Weichwald et al., 2015]. They produce weight maps that are needed to support this type of inference [Poldrack, 2011, Varoquaux et al., 2018]. These weight maps capture how brain regions relate to the outcome, *conditional on* the other regions, which is a key difference with respect to standard brain mapping based on mass univariate models. However, the weight maps produced by the common decoders come without statistical guarantees. Indeed, decoders optimize the quality of their prediction, but give no control on conditional feature importance. This is difficult due to the large number of covariates —voxels— as well as the severe multi-collinearity: voxel-level inference is untenable. On the other hand, given the spatial structure of the data, a spatial tolerance in the statistical control is natural, as in Gaussian random field theory used in standard analysis [Nichols, 2012].

Our first contribution is to formalize this spatial statistical control by introducing the δ-FWER, a control of false discoveries up to a spatial slack δ. This definition uncovers a fundamental trade-off between accuracy in the localization of the brain structures involved and statistical power: here we deliberately degrade spatial accuracy, acknowledging current concerns on statistical power in neuroimaging studies [Button et al., 2013, Noble et al., 2019].

Our second contribution is to study empirically the statistical control of four procedures computing decoding maps, ranging from thresholding procedures applied to SVR weights, to a dedicated decoding procedure, EnCluDL. Experiments show that the Thr-SVR procedure, thresholding SVR weights, fails to achieve useful statistical control. Exact permutation testing yields the expected statistical control but with very poor statistical power for all experimental settings we have studied. On the other hand, Adaptive Permutation Threshold SVR (Ada-SVR) [Gaonkar and Davatzikos, 2012], does not control the FWER as it should, though it exhibits a fair precision-recall curve in our semi-simulated experiments. This shows how difficult it is to identify a statistically valid threshold for SVR weight maps. This is due to the fact that under the null hypothesis, estimated weights are not distributed according to a fixed distribution —notably because of the dependency structure of the data— and more precisely, the variance of these distributions differs. Then, thresholding linear decoders (SVR, logistic regression) based on their estimated weights amplitudes is not a principled approach to control false discoveries.

EnCluDL uses a different decoding procedure to estimate the weight maps [Chevalier

et al., 2018], and as a result comes with theoretical statistical guarantees: it controls the δ-FWER for a predetermined tolerance parameter δ equal to the largest diameter of the clusters, assuming that the observed samples are i.i.d. and that the weight maps are homogeneous and sparse. The experiments show that, indeed, for i.i.d. scenarios, EnCluDL controls the δ-FWER for δ equal to the average radius of the clusters. Though, in some very high SNR or high sample size regimes, it might be necessary to take δ larger than the average radius (see Sec. 7.5). In practice, our choice of δ is conservative, and with current fMRI datasets, δ-FWER control holds for smaller δ, even in relatively large cohorts ($n = 1200$).

In our experiments, the spatial tolerance is around 1cm. Given that the definition of spatial location is blurred by inter-subject variability in group studies, this tolerance does not seem problematic. The method can thus be used for inference in cognitive neuroscience and population studies in psychiatry, neurology or epidemiology.

In addition, EnCluDL exhibits the best support recovery performance in the proposed semi-simulated experiments with fMRI data but also finds patterns with good face validity in more qualitative experiments plotted in Fig. 7. On the other hand, we also notice that EnCluDL tends to be over-conservative. Taking into account the difficulty of the problem and the fact that the convergence results are only asymptotic, we do not recommend using EnCluDL with $n < 100$.

In the present study, we have considered that the confounding variable effects have been removed during fMRI data preprocessing. However, it is still possible to include an additional confounding variable to the covariates before performing the inference. With regards to EnCluDL, we note that confounding variables should be handled separately from the clustered brain features.

Although it is not the main purpose of this study, we also checked the prediction performance of the decoders produced by each method. It is important to note that EnCluDL has been designed for the recovery of conditional statistical associations, not for prediction. In practice, the prediction performance is almost the same for SVR and Ada-SVR, and is slightly better than the one of EnCluDL (see Fig. 24). For prediction purpose, we recommend using *Fast Regularized Ensembles of Models* (FReM) [Hoyos-Idrobo et al., 2018], which is a stable and computationally efficient decoder with state-of-the-art prediction performance.

For pedagogical purpose, we have also considered a dataset where cross-sample independence is violated due to serial correlation, reproducing an experiment of Eklund et al. [2016]. The ensuing loss of statistical control underlines the importance of the i.i.d. hypothesis. Hence, EnCluDL should not be used to make inference from intra-subject dataset recorded over one session. With these warnings in mind, we think that EnCluDL can be used safely in neuroimaging context. Our code, implemented with Python 3, can be found on `https://github.com/ja-che/hidimstat` along with some examples.

We have not considered the method proposed by Nguyen et al. [2019] based on the Knockoff filters [Barber and Candès, 2015, Candès et al., 2018] that yet appear to be an appealing procedure, as it can only control the FDR. In this study we have focused on

δ-FWER control, and hence defer the analysis of FDR-controlling procedures to future work. Also, we have not benchmarked post-selection inference procedures [Lee et al., 2016, Berk et al., 2013], as we found them challenging to run in high dimensional settings and prone to numerical underflows.

Our empirical results clearly show that standard thresholding procedures, including classical permutation tests, are not reliable to infer regions importance on decoder maps, due to the high number of covariates. Since, in neuroimaging studies, these maps are used to give evidence on the brain regions that supports an outcome, it is crucial to use a procedure with statistical control on the brain maps. Our study shows that EnCluDL provides such a control.

Acknowledgements. Data were provided in part by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 826421, and has been supported by Labex DigiCosme (project ANR-11-LABEX-0045-DIGICOSME) operated by ANR as part of the program "Investissement d'Avenir" Idex Paris Saclay (ANR-11-IDEX-0003-02), as well as by the ANR-17-CE23-0011 Fast-Big project .

Ethics statement. No experiments on living beings were performed for this study. Hence, IRB approval was not necessary. The data that we used were acquired in original studies that had received approval by the original institution's IRB. All data were used accordingly to respective usage guidelines.

References

A. Abraham, F. Pedregosa, M. Eickenberg, P. Gervais, A. Mueller, J. Kossaifi, A. Gramfort, B. Thirion, and G. Varoquaux. Machine learning for neuroimaging with scikit-learn. *Frontiers in neuroinformatics*, 8:14, 2014. 13, 14

M. J. Anderson. Permutation tests for univariate or multivariate analysis of variance and regression. *Canadian journal of fisheries and aquatic sciences*, 58(3):626–639, 2001. 5

R. F. Barber and E. Candès. Controlling the false discovery rate via knockoffs. *Ann. Statist.*, 43(5):2055–2085, 10 2015. 7, 25

Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 57 (1):289–300, 1995. 4

R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao. Valid post-selection inference. *Ann. Statist.*, 41(2):802–837, 2013. 26

A. Bowring, F. Telschow, A. Schwartzman, and T. E. Nichols. Spatial confidence sets for raw effect size images. *NeuroImage*, 203:116187, 2019. 4

L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996. 6

P. Bühlmann. Statistical significance in high-dimensional linear models. *Bernoulli*, 19 (4):1212–1242, 09 2013. 6

K. S. Button, J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. J. Robinson, and M. R. Munafò. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14:365, 2013. 24

E. Candès, Y. Fan, L. Janson, and J. Lv. Panning for gold:'model-X' knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 80 (3):551–577, 2018. 7, 25

J.-A. Chevalier, J. Salmon, and B. Thirion. Statistical inference with ensemble of clustered desparsified lasso. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 638–646. Springer, 2018. 6, 7, 10, 11, 24

C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. 2, 15

B. Da Mota, V. Fritsch, G. Varoquaux, T. Banaschewski, G. J. Barker, A. L. W. Bokde, U. Bromberg, P. J. Conrod, J. Gallinat, H. Garavan, J.-L. Martinot, F. Nees, T. Paus, Z. Pausova, M. Rietschel, M. N. Smolka, A. Ströhle, V. Frouin, J.-B. Poline, and B. Thirion. Randomized parcellation based inference. *NeuroImage*, 89:203–215, 2014. 4

F. De Martino, G. Valente, N. Staeren, J. Ashburner, R. Goebel, and E. Formisano. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *Neuroimage*, 43(1):44–58, 2008. 2

O. Demirci, V. P. Clark, V. A. Magnotta, N. C. Andreasen, J. Lauriello, K. A. Kiehl, G. D. Pearlson, and V. D. Calhoun. A review of challenges in the use of fMRI for disease classification/characterization and a projection pursuit application from a multisite fMRI schizophrenia study. *Brain imaging and behavior*, 2(3):207–226, 2008. 2

R. Dezeure, P. Bühlmann, L. Meier, and N. Meinshausen. High-dimensional inference: Confidence intervals, p-values and R-Software hdi. *Statist. Sci.*, 30(4):533–558, 2015. 6, 33

O. J. Dunn. Multiple comparisons among means. *J. Amer. Statist. Assoc.*, 56(293): 52–64, 1961. 11

A. Eklund, T. Nichols, and H. Knutsson. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proc. Natl. Acad. Sci. U.S.A.*, 113 (28):7900–7905, 2016. 16, 25, 40

Y. Fan, N. Batmanghelich, C. M. Clark, C. Davatzikos, and Alzheimer's Disease Neuroimaging Initiative. Spatial patterns of brain atrophy in mci patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. *Neuroimage*, 39(4):1731–1743, 2008. 2

J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *Ann. Appl. Stat.*, 1(2):302–332, 2007. 6

K. J. Friston, A. P. Holmes, K. J. Worsley, J.-P. Poline, C. D. Frith, and R. S. J. Frackowiak. Statistical parametric maps in functional imaging: a general linear approach. *Human brain mapping*, 2(4):189–210, 1994. 13

B. Gaonkar and C. Davatzikos. Deriving statistical significance maps for svm based image classification and group comparisons. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 723–730. Springer, 2012. 5, 9, 18, 24

J. R. Gimenez and J. Zou. Discovering conditionally salient features with statistical guarantees. *International Conference on Machine Learning*, pages 2290–2298, 2019. 4

M. F. Glasser, T. S. Coalson, E. C. Robinson, C. D. Hacker, J. Harwell, E. Yacoub, K. Ugurbil, J. Andersson, C. F. Beckmann, M. Jenkinson, S. M. Smith, and D. C. Van Essen. A multi-modal parcellation of human cerebral cortex. *Nature*, 536:171–178, 2016. 4

A. Gramfort, G. Varoquaux, and B. Thirion. Beyond brain reading: randomized sparsity and clustering to simultaneously predict and identify. In *Machine Learning and Interpretation in Neuroimaging*, pages 9–16. Springer, 2012. 6

A. Gramfort, B. Thirion, and G. Varoquaux. Identifying predictive regions from fMRI with TV-L1 prior. In *2013 International Workshop on Pattern Recognition in Neuroimaging*, pages 17–20. IEEE, 2013. 2

S. Haufe, Frank Meinecke, Kai G., S. Dähne, J.-D. Haynes, B. Blankertz, and F. Bießmann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage*, 87:96–110, 2014. 2

J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001. 2

J.-D. Haynes and G. Rees. Neuroimaging: decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7(7):523, 2006. 2

T. K. Ho. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(8):832–844, 1998. 6

Y. Hochberg and A. C. Tamhane. *Multiple comparison procedures*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 1987. 3, 4, 8

A. Hoyos-Idrobo, G. Varoquaux, Y. Schwartz, and B. Thirion. Frem–scalable and stable decoding with fast regularized ensemble of models. *NeuroImage*, 180:160–172, 2018. 6, 11, 12, 25

L. Janson and W. Su. Familywise error rate control via knockoffs. *Electron. J. Stat.*, 10 (1):960–975, 2016. 7

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.*, 15:2869–2909, 2014. 6, 33

L. I. Kuncheva and J. J. Rodríguez. Classifier ensembles for fMRI data analysis: an experiment. *Magnetic resonance imaging*, 28(4):583–593, 2010. 6

L. I. Kuncheva, J. J. Rodríguez, C. O. Plumpton, D. E. J. Linden, and S. J. Johnston. Random subspace ensembles for fMRI classification. *IEEE Trans. Med. Imaging*, 29 (2):531–542, 2010. 6

J. Lee, D. Sun, Y. Sun, and J. Taylor. Exact post-selection inference, with application to the lasso. *Ann. Statist.*, 44(3):907–927, 2016. 26

S. Lee, S. Halder, A. Kübler, N. Birbaumer, and R. Sitaram. Effective functional mapping of fmri data with support-vector machines. *Human brain mapping*, 31(10):1502–1511, 2010. 2

N. Meinshausen, L. Meier, and P. Bühlmann. P-values for high-dimensional regression. *J. Amer. Statist. Assoc.*, 104(488):1671–1681, 2009. 6, 11, 34

J. Mourao-Miranda, A. L. W. Bokde, C. Born, H. Hampel, and M. Stetter. Classifying brain states and determining the discriminating activation patterns: support vector machine on functional MRI data. *NeuroImage*, 28(4):980–995, 2005. 2

T.-B. Nguyen, J.-A. Chevalier, and B. Thirion. Ecko: Ensemble of clustered knockoffs for robust multivariate inference on fMRI data. In *International Conference on Information Processing in Medical Imaging*, pages 454–466. Springer, 2019. 4, 7, 25

T. E. Nichols. Multiple testing corrections, nonparametric methods, and random field theory. *Neuroimage*, 62:811, 2012. 24

S. Noble, D. Scheinost, and R. Constable. Cluster failure or power failure? evaluating sensitivity in cluster-level inference. *NeuroImage*, 209:116468, 12 2019. 24

K. A. Norman, S. M. Polyn, G. J. Detre, and J. V. Haxby. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in cognitive sciences*, 10(9):424–430, 2006. 2

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011. 13

F. Pereira, T. Mitchell, and M. Botvinick. Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage*, 45(1):S199–S209, 2009. 2, 5

A. L. Pinho, A. Amadon, T. Ruest, M. Fabre, E. Dohmatob, I. Denghien, C. Ginisty, S. Becuwe-Desmidt, S. Roger, L. Laurier, et al. Individual brain charting, a high-resolution fMRI dataset for cognitive mapping. *Scientific data*, 5:180105, 2018. 14

R. A. Poldrack. Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. *Neuron*, 72(5):692–697, 2011. 2, 24

R. A. Poldrack, J. A. Mumford, and T. E. Nichols. *Handbook of functional MRI data analysis*. Cambridge University Press, 2011. 2

A. K. Rehme, L. J. Volz, D.-L. Feis, I. Bomilcar-Focke, T. Liebig, S. B. Eickhoff, G. R. Fink, and C. Grefkes. Identifying neuroimaging markers of motor disability in acute stroke by machine learning techniques. *Cerebral cortex*, 25(9):3046–3056, 2015. 2

A. Rizk-Jackson, D. Stoffers, S. Sheldon, J. Kuperman, A. Dale, J. Goldstein, J. Corey-Bloom, R. A. Poldrack, and A. R. Aron. Evaluating imaging biomarkers for neurodegeneration in pre-symptomatic Huntington's disease using machine learning techniques. *Neuroimage*, 56(2):788–796, 2011. 2, 5

J. R. Sato, R. Basilio, F. F. Paiva, G. J. Garrido, I. E. Bramati, P. Bado, F. Tovar-Moll, R. Zahn, and J. Moll. Real-time fmri pattern decoding and neurofeedback using friend: an fsl-integrated bci toolbox. *PLoS One*, 8(12):e81658, 2013. 2

Y. Schwartz, B. Thirion, and G. Varoquaux. Mapping coginitive ontologies to and from the brain. In *Advances in neural information processing systems*, pages 1673–1681, 2013. 2

A. Schwartzman, R. F. Dougherty, J. Lee, D. Ghahremani, and J. E. Taylor. Empirical null and false discovery rate analysis in neuroimaging. *Neuroimage*, 44(1):71–82, 2009. 9

S. M. Smith and T. E. Nichols. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage*, 44(1):83–98, 2009. 4

A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Stat. Comput.*, 14(3):199–222, 2004. 2

B. Thirion, G. Varoquaux, E. Dohmatob, and J.-B. Poline. Which fMRI clustering gives good brain parcellations? *Frontiers in Neuroscience*, 8:167, 2014. 6, 11

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 58(1):267–288, 1996. 6

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42(3):1166–1202, 2014. 6, 11, 12, 40

S. Van der Walt, S. C. Colbert, and G. Varoquaux. The numpy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22–30, 2011. 13

D. C. Van Essen, K. Ugurbil, E. J. Auerbach, D. M. Barch, T. E. J. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S. W. Curtiss, S. Della Penna, D. A. Feinberg, M. F. Glasser, N. Harel, A. C. Heath, L. J. Larson-Prior, D. S. Marcus, G. Michalareas, S. Moeller, R. Oostenveld, S. E. Petersen, F. W. Prior, B. L. Schlaggar, S. M. Smith, A. Z. Snyder, J. Xu, and E. Yacoub. The Human Connectome Project: a data acquisition perspective. *Neuroimage*, 62(4):2222–2231, 2012. 14

G. Varoquaux and B. Thirion. How machine learning is shaping cognitive neuroimaging. *GigaScience*, 3(1):28, 2014. 2

G. Varoquaux, A. Gramfort, and B. Thirion. Small-sample brain mapping: sparse recovery on spatially correlated designs with randomization and clustering. In *International Conference on Machine Learning*, 2012. 6, 11

G. Varoquaux, Y. Schwartz, R. A. Poldrack, B. Gauthier, D. Bzdok, J.-B. Poline, and B. Thirion. Atlases of cognition with large-scale human brain mapping. *PLoS computational biology*, 14(11):e1006565, 2018. 2, 24

P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, I. Polat, Y. Feng, E. W. Moore, J. Vand erPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1. 0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 2020. 13

M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Trans. Image Process.*, 55 (5):2183–2202, 2009. 5

Y. Wang, J. Zheng, S. Zhang, X. Duan, and H. Chen. Randomized structural sparsity via constrained block subsampling for improved sensitivity of discriminative voxel identification. *Neuroimage*, 117:170–183, 2015. 6

L. Wasserman and K. Roeder. High-dimensional variable selection. *Ann. Statist.*, 37 (5A):2178–2201, 2009. 6

S. Weichwald, T. Meyer, O. Özdenizci, B. Schölkopf, T. Ball, and M. Grosse-Wentrup. Causal interpretation rules for encoding and decoding models in neuroimaging. *Neuroimage*, 110:48–59, 2015. 2, 3, 13, 20, 24

P. H. Westfall and S. S. Young. *Resampling-based multiple testing: Examples and methods for p-value adjustment*, volume 279. John Wiley & Sons, 1993. 5, 9, 13

C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 76(1):217–242, 2014. 6, 11, 32, 33, 40

Z.-H. Zhou. *Ensemble methods: foundations and algorithms*. Chapman and Hall/CRC, 2012. 6

7 Appendix

7.1 Desparsified Lasso

Additional notation. For a matrix $\mathbf{X}$, $\mathbf{X}_{i,\cdot}$ refers to the i-th row and $\mathbf{X}_{\cdot,j}$ to the j-th column, $\mathbf{X}_{i,j}$ refers to the element (i,j), and $\mathbf{X}^{(-j)}$ refers to the matrix $\mathbf{X}$ without the j-th column. $\mathbf{X}^\dagger$ denotes the Moore-Penrose pseudo-inverse of $\mathbf{X}$.

Small-dimension insight. The Desparsified Lasso procedure, introduced by Zhang and Zhang [2014] extends the Ordinary Least Squares (OLS) procedure to $n < p$ cases. Let us first recall the standard OLS framework ($n > p$). Starting from model (1), let us define $\mathbf{z}_j \in \mathbb{R}^n$ the residual of the OLS regression of $\mathbf{X}_{\cdot,j}$ versus $\mathbf{X}^{(-j)}$ given by:

$$\mathbf{z}_j = \mathbf{X}_{\cdot,j} - \mathbf{X}^{(-j)}\hat{\mathbf{w}}^{(-j)} \quad , \tag{14}$$

where $\hat{\mathbf{w}}^{(-j)}$ refers to the estimator of the OLS regression of $\mathbf{X}_{\cdot,j}$ versus $\mathbf{X}^{(-j)}$. In particular, $\mathbf{z}_j^\top \mathbf{X}_{\cdot,k} = 0$ for all $k \in [p] \setminus \{j\}$. In this setting, we also have the following result:

Proposition 7.1. *If $n > p$ and $rank(\mathbf{X}) = p$, then, for all $j \in [p]$:*

$$\hat{\mathbf{w}}_j^{\mathrm{OLS}} = \frac{\mathbf{z}_j^\top \mathbf{y}}{\mathbf{z}_j^\top \mathbf{X}_{\cdot,j}} \quad , \tag{15}$$

where $\hat{\mathbf{w}}^{\mathrm{OLS}}$ is the parameter vector estimates obtained from the OLS regression of $\mathbf{y}$ against $\mathbf{X}$.

Desparsified Lasso. In this setting, it is not possible to construct a non-zero vector family $\{\mathbf{z}_j, j \in [p]\}$ (*i.e.,* a family verifying $\mathbf{z}_j \neq \mathbf{0}$ for all $j \in [p]$), such that $\mathbf{z}_j^\top \mathbf{X}_{\cdot,k} = 0$ for all $k \neq j$. The idea proposed by Zhang and Zhang [2014] is to construct a family $\{\mathbf{z}_j, j \in [p]\}$ which would play the same role as the residual of the OLS regression of $\mathbf{X}_{\cdot,j}$ versus $\mathbf{X}^{(-j)}$ in (14) but relaxing (slightly) the constraint $\mathbf{z}_j^\top \mathbf{X}_{\cdot,k} = 0$. To do so, instead of computing $\{\mathbf{z}_j, j \in [p]\}$ by OLS regression, they proposed to take the residual of the Lasso regressions[1] of $\mathbf{X}_{\cdot,j}$ against $\mathbf{X}^{(-j)}$. Then, from (1), one can derive the following:

$$\frac{\mathbf{z}_j^\top \mathbf{y}}{\mathbf{z}_j^\top \mathbf{X}_{\cdot,j}} = \mathbf{w}_j^* + \frac{\mathbf{z}_j^\top \boldsymbol{\varepsilon}}{\mathbf{z}_j^\top \mathbf{X}_{\cdot,j}} + \sum_{k \neq j} \frac{\mathbf{z}_j^\top \mathbf{X}_{\cdot,k} \mathbf{w}_k^*}{\mathbf{z}_j^\top \mathbf{X}_{\cdot,j}} \ . \tag{16}$$

Noticing that the second term in (16) is a noise term and plugging in an initial estimator $\hat{\mathbf{w}}^{(\text{init})}$ of $\mathbf{w}^*$ in the third term —a standard choice being the Lasso— they propose the following estimator:

$$\hat{\mathbf{w}}_j = \frac{\mathbf{z}_j^\top \mathbf{y}}{\mathbf{z}_j^\top \mathbf{X}_{\cdot,j}} - \sum_{k \neq j} \frac{\mathbf{z}_j^\top \mathbf{X}_{\cdot,k} \hat{\mathbf{w}}_k^{(\text{init})}}{\mathbf{z}_j^\top \mathbf{X}_{\cdot,j}} \ . \tag{17}$$

Here, one can notice that (17) generalizes (15) to $n < p$. Then, from (16) and (17) one can derive:

$$\sigma_\varepsilon^{-1}(\hat{\mathbf{w}}_j - \mathbf{w}_j^*) = \underbrace{\sigma_\varepsilon^{-1} \frac{\mathbf{z}_j^\top \boldsymbol{\varepsilon}}{\mathbf{z}_j^\top \mathbf{X}_{\cdot,j}}}_{\boldsymbol{\eta}_j} + \underbrace{\sigma_\varepsilon^{-1} \sum_{k \neq j} \frac{\mathbf{z}_j^\top \mathbf{X}_{\cdot,k}}{\mathbf{z}_j^\top \mathbf{X}_{\cdot,j}} (\mathbf{w}_k^* - \hat{\mathbf{w}}_k^{(\text{init})})}_{\boldsymbol{\mu}_j} \ . \tag{18}$$

This yields:

$$\sigma_\varepsilon^{-1}(\hat{\mathbf{w}} - \mathbf{w}^*) = \boldsymbol{\eta} + \boldsymbol{\mu}, \qquad \boldsymbol{\eta} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Omega}) \ , \tag{19}$$

where:

$$\boldsymbol{\Omega}_{jk} = \frac{\mathbf{z}_j^\top \mathbf{z}_k}{(\mathbf{z}_j^\top \mathbf{X}_{\cdot,j})(\mathbf{z}_k^\top \mathbf{X}_{\cdot,k})} \ . \tag{20}$$

Asymptomatically and under some sparsity assumptions (one can refer to [Dezeure et al., 2015] for more details), one can neglect the last term $\boldsymbol{\mu}$ and obtain:

$$\sigma_\varepsilon^{-1}(\boldsymbol{\Omega}_{jj})^{-1/2}(\hat{\mathbf{w}}_j - \mathbf{w}_j^*) \sim \mathcal{N}(0, 1) \ . \tag{21}$$

From (21), one can compute the confidence intervals and p-values of the coefficients of the estimated weight map. Note that similar estimators have been derived in parallel in Javanmard and Montanari [2014].

[1] From our analysis, taking λ_j, the regularization parameter used in the Lasso regression of $\mathbf{X}_{\cdot,j}$ against $\mathbf{X}^{(-j)}$, equal to $0.01 \times \max_{k \in [p] \setminus \{j\}} |\mathbf{X}_{\cdot,j}^\top \mathbf{X}_{\cdot,k}|/n$ is appropriate to compute $\mathbf{z}_j$. Empirically, it results in a more conservative solution than the one proposed by Zhang and Zhang [2014] but it avoids doing computationally expensive grid-search.

7.2 Adaptive quantile aggregation of p-values

For the j-th voxel, we have a vector $(p_j^{(b)})_{b\in[B]}$ of p-values, with one p-value computed for each of the B clusterings. Then, the final p-value of the j-th feature is given by the adaptive quantile aggregation, as proposed by Meinshausen et al. [2009]:

$$p_j = \min\left\{(1-\log(\gamma_{\min}))\inf_{\gamma\in(\gamma_{\min},1)}\left(\gamma\text{-quantile}\left\{\frac{p_j^{(b)}}{\gamma}; b\in[B]\right\}\right),\ 1\right\}\ ,$$

where we have taken $\gamma_{\min} = 0.20$ in our experiments. Taking a value of $\gamma_{\min}$ not too small (*e.g.,* $\gamma_{\min} \geq 0.20$) ensures that the discovered sources have received small p-values many times (*e.g.,* at least for $B/5$ different choices of clustering).

7.3 Empirical analysis of data structure impact

In this section, we propose two simulations to gain more insight concerning the assumptions about data structure that are necessary for Desparsified Lasso and EnCluDL to have power. More precisely, we investigate up to which level of correlation two correlated predictive features (having non-zero weight) are both identified. Indeed, when two predictive features are highly correlated, there is a risk that the inference procedure only detects one of the two.

The first simulation has modest data dimension, which corresponds to that of data after clustering. We use it to analyze the behavior of Desparsified Lasso. The second simulation has a 2D structure with larger data dimension, it introduces short- and long-range correlation structure, it is used to study EnCluDL.

First simulation: approximating the clustered data setting. In this simulation we set $n = 100$ and $p = 500$. We construct the design matrix $\mathbf{X}$ such that features are normally distributed and the first two features have a correlation equal to parameter ρ, while all the other features are independent. The weight $\mathbf{w}^*$ is such that $\mathbf{w}^*_j = 1$ for $1 \leq j \leq 10$ and $\mathbf{w}^*_j = 0$ otherwise. We also set $\sigma_\varepsilon = 1$ giving approximately $\text{SNR}_y = 12$ close to the SNR estimated in real fMRI datasets.

To check the ability of Desparsified Lasso to identify two correlated features, we compare the smallest z-score of the first two first features ("correlated features") with the smallest z-score of the two following features ("control features") for different value of $\rho \in (0, 1)$. While the minimum z-score of the control features should not vary significantly and corresponds to a control value, the minimum z-score of the two correlated features should decrease towards 0 when ρ increases to 1. Also, we look at the z-score of a random non-predictive feature ("random null feature") to get insight about the z-score threshold value to declare a feature significant.

First simulation results. In Fig. 10, we give the results for the first simulation. When the correlation of the two correlated features increases, their identification using the Desparsified Lasso procedure becomes harder. In this experiment, we observe that

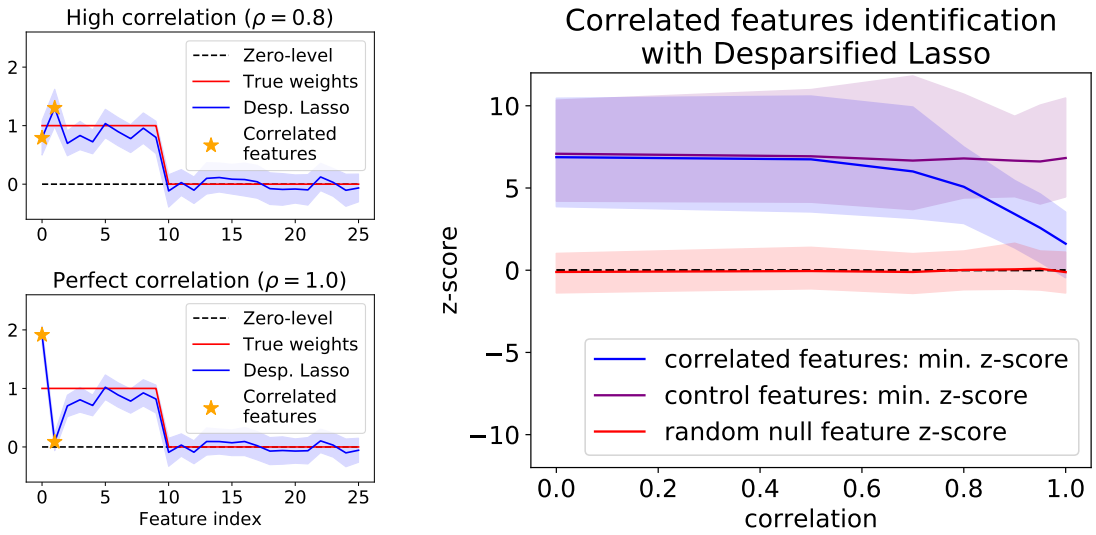


Figure 10: **Impact of correlation when trying to identify two correlated features.** Left: We plot the Desparsified Lasso estimator and its 95% confidence intervals. The correlation between the first two features is set to ρ, while the other features are uncorrelated. The higher ρ the harder it is to identify each of the two correlated features. For $\rho = 1.0$, it is impossible, while for $\rho = 0.8$, the identification of both features is successful. Right: Quantitative summary of the simulations. When the correlation increases the minimum z-score of the two first features ("correlated features") decreases (90% confidence intervals also displayed). The correlation between the two following features ("control features") remains equal to zero, thus the minimum z-score of these features is used as a control value that should not vary significantly. Also we plot the z-score of a random non-predictive feature ("random null feature"). We observe that for a correlation lower than 0.8 the deviation is limited and it is possible to identify the two correlated variables. For a correlation larger than 0.9 the deviation is massive and it becomes impossible to recover the two correlated variables.

below a correlation of 0.8, Desparsified Lasso can identify accurately the two correlated variables. However, above a correlation of 0.9, Desparsified Lasso might fail to recover the both correlated variables.

Second simulation: 2D data structure. The simulation we consider has a 2D data structure. It aims at approximating the short- and long-range correlation structure that can be observed in fMRI data (see Sec. 7.4). The feature space considered is a square with edge length $H = 40$, then $p = H^2 = 1\,600$ features and we took $n = 100$ samples. To construct $\mathbf{w}^*$, we define a 2D weight map $\tilde{\mathbf{w}}^*$ of size $H \times H$ with four active regions then we flatten $\tilde{\mathbf{w}}^*$ in a vector $\mathbf{w}^*$ of size p. Each active region is a small square of width $h = 4$, leading to support of size $4 \times h^2 = 64$. The four active regions are located in the corners of the weight map. The true weight map is represented in Fig. 11. To construct

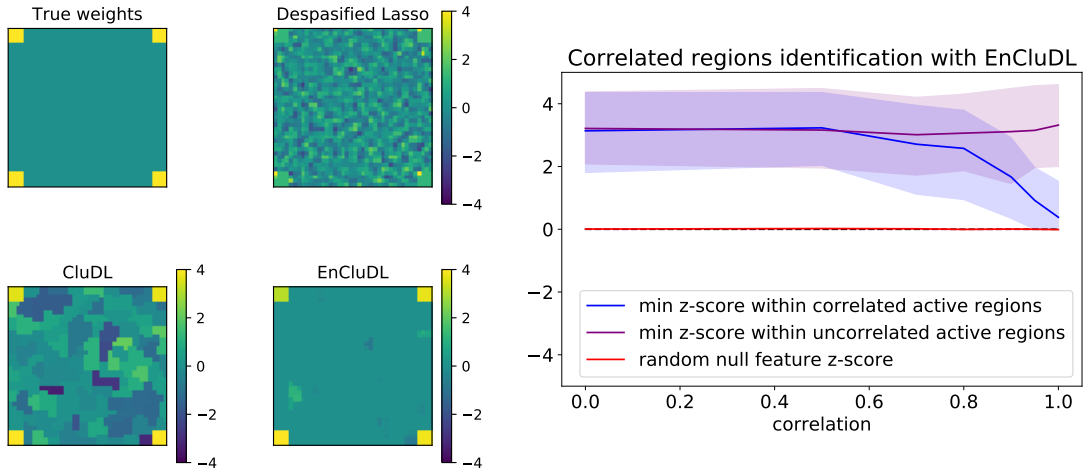


Figure 11: **Impact of correlation when trying to identify two correlated regions.** Left: True weight map, and z-scores estimated by Desparsified Lasso, CluDL and EnCluDL, obtained for $\rho = 0.9$. Desparsified Lasso cannot handle the extreme short-range correlation that occurs within each predictive region and only identifies one feature in each. CluDL and EnCluCL benefit from the clustering, as they identify all the features for every predictive regions. We can also observe that EnCluDL improves upon CluDL thanks to the smoothing effect produced by ensembling. Focusing on the EnCluDL solution, we can see that the z-score of the upper left active region is a bit lower than for the other active regions. This is due to the high correlation between the upper left and bottom right regions. Right: Summary of the results of the second simulation. When the correlation increases the minimum z-score within the correlated active regions decreases. The minimum z-score between the two uncorrelated regions is used as a control. We also plot the z-score of a random non-predictive feature, we notice that due to the ensembling step of EnCluDL, the empirical confidence intervals are much thinner than in Fig. 10. We observe that for a correlation lower than 0.8 the deviation is limited and it is possible to identify the two correlated predictive regions. For a correlation larger than 0.9 the deviation becomes large and recovering the two correlated regions becomes impossible.

the design matrix $\mathbf{X}$, we first construct a 2D matrix $\tilde{\mathbf{M}}$ by drawing p random normal vectors of size n that are spatially smoothed with a 2D Gaussian filter (the smoothing is only made in the feature space for each sample independently, the samples are not mixed and remain independent). We flatten the vectors to go from $\tilde{\mathbf{M}}$ of size $n \times H \times H$ to $\mathbf{M}$ of size $n \times p$. The spatial smoothing enforces a 2D structure on the data. Then, we further modify $\mathbf{M}$ such that (i) all the features of an active region are perfectly correlated and (ii) two of the four active regions are correlated at a given value $\rho \in (0, 1)$, the two other active regions being unmodified (hence uncorrelated). The first transformation aims at showing that the clustering is useful to handle the short-range correlation that might be very high for fMRI data (see Sec. 7.4). The second transformation aims at

testing whether EnCluDL can recover two correlated predictive regions; this is notably desirable in the case of long-range correlation (*e.g.,* two contralateral brain regions). The two uncorrelated regions are used to provide control values. With these transformations we obtain the design matrix $\mathbf{X}$. In Sec. 7.4, the two active regions that are correlated are located in the upper left corner and in the bottom right corner while the other two are uncorrelated. Finally, we also set $\sigma_\varepsilon = 10$, to approximately get $\mathrm{SNR}_y = 4$.

To check the ability of EnCluDL to identify two correlated regions, we compare the smallest z-score of the features that belong to one of the correlated regions with the smallest z-score of the features that belong to the uncorrelated active regions; we analyze the results for several values of $\rho \in (0, 1)$. To understand the effect of the clustering and ensembling, we compare Desparsified Lasso, CluDL and EnCluDL solutions qualitatively. Since the features that belong to the same active region are perfectly correlated, we expect that Desparsified Lasso identifies only one feature per region at best. We also report the z-score of a random non-predictive feature.

Second simulation results. In Fig. 11, we give the results for the second simulation. Clustering turns out to be crucial to produce valid statistical inference solution in the presence of extreme short-range correlation. Additionally, we show that when the correlation of the two correlated active regions increases, their identification using EnCluDL becomes harder. In this experiment, we observe that below a correlation of 0.8, EnCluDL can identify accurately the two correlated regions. However, above a correlation of 0.9, EnCluDL generally fails to recover the two correlated regions.

7.4 fMRI data structure

In Sec. 7.3, we have shown that one may encounter multicollinearity issues. It is thus necessary to analyze the correlation structure of actual fMRI data.

In Fig. 12, we study the correlation observed in the HCP900 Emotion task data. Considering correlation between random voxels, then neighboring voxels, we can see that the correlation is much higher in the case of neighboring voxel. Notably, the median correlation between two random voxels is 0.1 while the median correlation between two neighboring voxels is above 0.8, and often larger than 0.9. We have shown in Sec. 7.3, that Desparsified Lasso may fail to detect two features when they are so strongly correlated.

Correlation histograms after clustering the data as shown in Fig. 12. For example, taking $C = 500$ clusters, the median correlation between two random clusters is 0.3 while it is 0.7 for two neighboring clusters. Inter-cluster correlation always remains below 0.85 and almost always below 0.8. In practice, we have shown in Sec. 7.3 that Desparsified Lasso can handle scenarios where features have correlation lower than 0.8.

7.5 Estimating δ for which EnCluDL controls the δ-FWER

In Sec. 3.5, we recommend using δ, in regular brain imaging settings with (12):

$$\delta_0 = \left(\frac{p}{2C}\right)^{1/3} ,$$

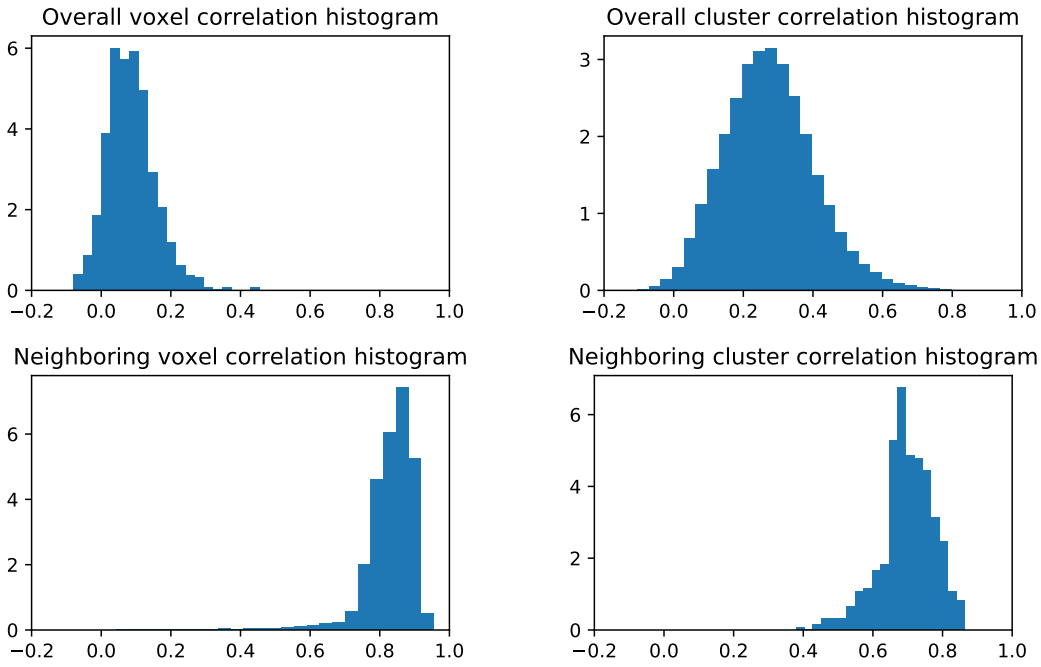


Figure 12: **Data structure in HCP900 emotion task.** Left: Correlation histogram of the fMRI data at voxel level. The correlation between two random voxels is quite low, a typical value being around 0.1. However, when looking at neighboring voxels, we observe that the correlation is often higher then 0.9. This exhibits the short- and long-range correlation structure but also suggests that raw Desparsified Lasso would not be adapted to this setting. Right: Correlation histogram of the clustered data for $C = 500$. The correlation between two random clusters is around 0.3, while the correlation between two neighboring clusters is around 0.7 and almost always below 0.8. Then, thanks to clustering, highly correlated voxels are aggregated into groups and Desparsified Lasso is adapted to this setting.

δ_0 being a distance in voxel unit close to the average radius of the clusters used in EnCluDL. However, when the setting is particularly favorable for inference, *i.e.,* if $\log(n)/C$ is large or σ_ε is small, the choice of δ given by (12) may be over-optimistic and we might need to correct this formula. We have found empirically that a suitable multiplicative factor, denoted by $\tau > 0$, that could be used to correct δ_0 is given by:

$$\tau = -45 \log\left(\frac{\sigma_\varepsilon}{\mathrm{std}(\mathbf{y})}\right) \frac{\log(n)}{C} \quad , \tag{22}$$

where σ_ε is the standard deviation of the noise ε. In practice σ_ε has to be estimated; in the fMRI datasets we studied, estimates of $\frac{\sigma_\varepsilon}{\mathrm{std}(\mathbf{y})}$ were close to 0.1. However, given the heuristic derivation of this quantity and the uncertainty about the value of τ, we do not recommend correcting δ_0 with a factor lower than 1 as it could lead to a dramatic under estimation of the valid δ. Then, the final formula to compute the δ such that δ-FWER

control is ensured, is:

$$\delta^* = \max(1, \tau)\ \delta_0\ . \tag{23}$$

Note that the formula given by (12) and even (23) are not bullet proof but rather give reasonable estimates of δ.

7.6 Cluster size analysis

In Sec. 3.5, we have proposed a formula to compute a valid spatial tolerance parameter δ_0. In Fig. 13, we show that δ_0 is close but slightly lower than the average cluster radius. Also, one can notice that taking a larger number of clusters, the size of the clusters is smaller. As a consequence, the statistical control is valid for a lower spatial tolerance. Finally, by looking at the shape of the distribution of the cluster radius, we observe that there are only few large clusters.

In general δ_0 is a suitable choice, however when the setting is particularly favorable for inference, the mixing effect produced by ensembling might not be sufficient and voxels far (further than δ_0) from the support might be discovered. This effect can be explained by the detection of large clusters that are overlapping the support and the null region.

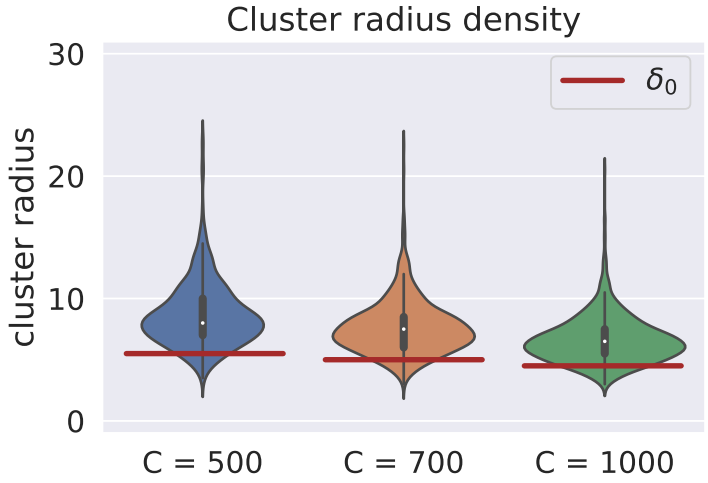


Figure 13: **Comparing δ_0 with the distribution of the cluster radius as a funtion of C.** By taking a larger number of clusters, we decrease the size of the clusters. The statistical control is thus valid for a smaller spatial tolerance. Comparing the distribution of the cluster radius with the recommended choice of spatial tolerance parameter δ_0, we observe that δ_0 is a bit lower than the empirical average cluster radius. Finally, we observe that few clusters are much wider than the others, this may occasionally lead to false discoveries far from the support in high SNR scenarios.

7.7 Illustrating spatial tolerance on real brain geometry

In Fig. 14, we display a brain pattern with spatial tolerance in the case of the HCP data.

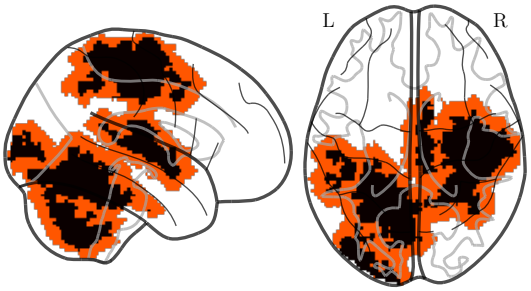


Figure 14: **Expanding HCP maps by 6 voxels.** The black-colored voxels represent the positive weights of the reference map constructed in Sec. 4.2. The red-colored voxels are the δ-dilation of the previous map where $\delta = 6$ voxels, *i.e.,* the tolerance we have taken in all experiments. Then, δ-FWER controls the false discoveries made outside of the colored voxels (see also Sec. 3.1).

7.8 Statistical control under the global null with autocorrelated data

Experiment. In this experiment, we study how the different procedures control the FWER when the data are temporally autocorrelated; hence violating the i.i.d. assumption. Notably, this is the case if the data correspond to fMRI signal recordings of one given subject during an acquisition. We consider data from the HCP900 resting-state fMRI dataset described in Sec. 4.1 with full samples ($n = 1200$). The design matrix $\mathbf{X}$ contains the 15-minutes fMRI signal records. As in Eklund et al. [2016], we construct $\mathbf{y}$ such that it corresponds to two activity paradigms: block or event responses, with several frequencies: 10s on/off, 20s on/off, 30s on/off, 2s-activation/6s-rest, 4s-activation/8s-rest. Thus, $\mathbf{y}$ is temporally autocorrelated. In these simulations $\mathbf{w}^* = \mathbf{0}$ so the δ-FWER and the classical FWER are identical. To better assess the impact of correlation, we also generate $\mathbf{y}$ as an i.i.d. —uncorrelated— Bernoulli or standard Gaussian random variable (here again $\mathbf{w}^* = \mathbf{0}$), breaking spurious correlations between $\mathbf{X}$ and $\mathbf{y}$. These two cases enable to check if the procedures still control the FWER at the targeted nominal level on this dataset under the i.i.d. hypothesis. For each kind of response, we repeat the experiment 100 times, using data from 100 different subjects.

Results. we now report the results of the experiment. In Fig. 15, we observe that for all the fictitious block response paradigms, for every procedure, the empirical FWER exceeds the targeted nominal level (10%), as one would expect. This result is not surprising since independence across samples is a key assumption for a valid statistical inference with any of the four procedures. Notably, concerning EnCluDL, Desparsified Lasso needs the i.i.d. hypothesis [Zhang and Zhang, 2014, van de Geer et al., 2014] to produce valid confidence intervals or p-values. This assumption is not verified for the block or event

response paradigms due to the temporal dependency in the data. However, when the target $\mathbf{y}$ is i.i.d. —*i.e.,* without temporal dependency (Bernoulli or Gaussian random responses)— the FWER is controlled (except for Thr-SVR). Indeed, the model is no longer confounded by the correlation structure underlying the data.

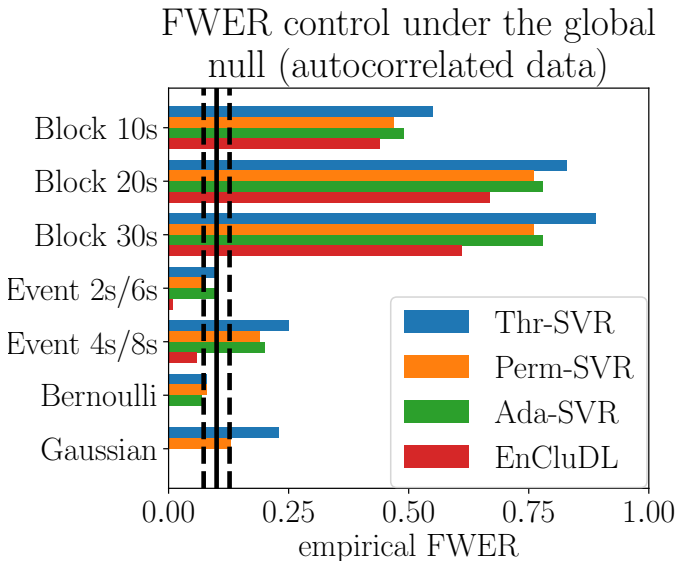
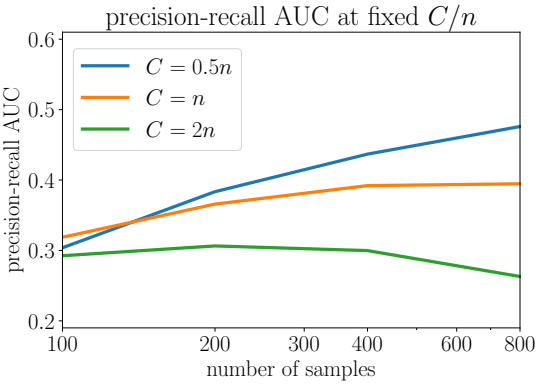


Figure 15: **FWER control under the global null with autocorrelated data.** The results of the experiment with correlated data under the global null, described in Sec. 7.8, show that, when the data are temporally autocorrelated, all the procedures fail to control the FWER. Indeed, for all the fictitious block response paradigms, the empirical FWER exceeds the targeted nominal level of 10% for every procedure. This result is not surprising as the procedures control the δ-FWER under the hypothesis that the samples are i.i.d.; this is not the case for the block or event response paradigms. However, when the fictitious response breaks the temporal dependency (binary or Gaussian random responses), the i.i.d. hypothesis is met and the FWER is empirically well controlled except for the Thr-SVR procedure.

7.9 Influence of the C/n ratio on the recovery property of EnCluDL

When using EnCluDL, the number C of clusters is an arbitrary parameter. We proposed some default choice in Sec. 4.5, yet intuitively, C should adapt to the amount of data available: larger samples size lead to better estimation, allowing refined localization, hence higher C. In Fig. 16, we show on semi-simulated data that for $C \in [n/2, n]$, C/n being fixed, the precision-recall AUC on real data does not depend on n, suggesting to chose C proportional to n.

Figure 16: **Influence of the C/n ratio on the precision-recall AUC.** The results of the experiment described in Sec. 4.5 show that the precision-recall AUC depends almost linearly on $\log(C/n)$ except when C is critically low creating very wide clusters and deteriorating the precision-recall curve. This limit depends on the physical properties of the problem; here, C should not be lower than 100. Keeping this limit in mind, we advise taking $C \in [n/2, n]$ to recover most of the predictive regions.



7.10 Statistical control with known ground truth: additional plots

In this section, we provide additional experimental results to assess the detection accuracy of the multivariate estimators, to complement the results in Sec. 4.2. Fig. 17 shows additional precision-recall curves, obtained for different values of n: these different settings preserve the relative performance of the methods, while larger n results in better curves. However, we do not recommend running such analysis with $n < 100$, since the estimation problem is hard and statistical guarantees only hold in asymptotic regime. Fig. 18 and Fig. 19 display the performance of the methods in terms of δ-FWER control and precision-recall curves on semi-simulated data where $\mathbf{y}$ is binary. This induces a violation of the EnCluDL model that reduces its performance in terms of δ precision-recall. Yet, unlike Ada-SVR, it still controls the δ-FWER accurately.

7.11 Face validity on HCP dataset

In Fig. 23, we plot the results for five tasks taken from the HCP dataset, besides of the two described in Sec. 4.6. For all methods, the statistical maps are thresholded such that the δ-FWER stays lower than 10% for $\delta = 12\,\text{mm}$. Qualitatively, EnCluDL discovers the most plausible patterns, Ada-SVR often makes dubious discoveries, patterns are too wide and implausible, while the two other methods exhibit a very weak statistical power. As discussed in the main person, Univ-OLS provides complementary results that highlight marginal association between the data and the target.

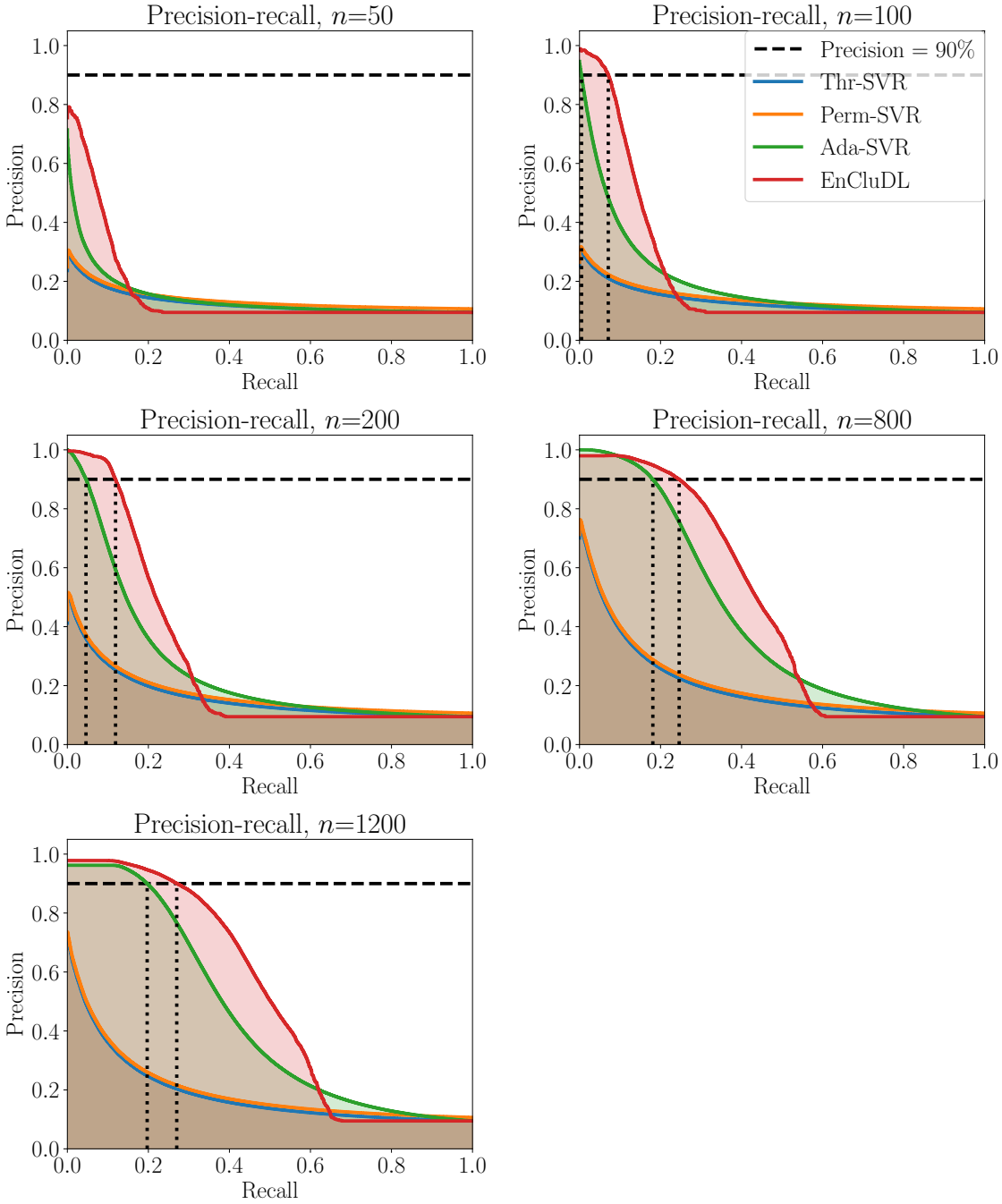


Figure 17: **Precision-recall curves on semi-simulated data with continuous response vector.** The results of the experiment described in Sec. 4.2 show that EnCluDL has the best performance in terms of precision-recall curve.

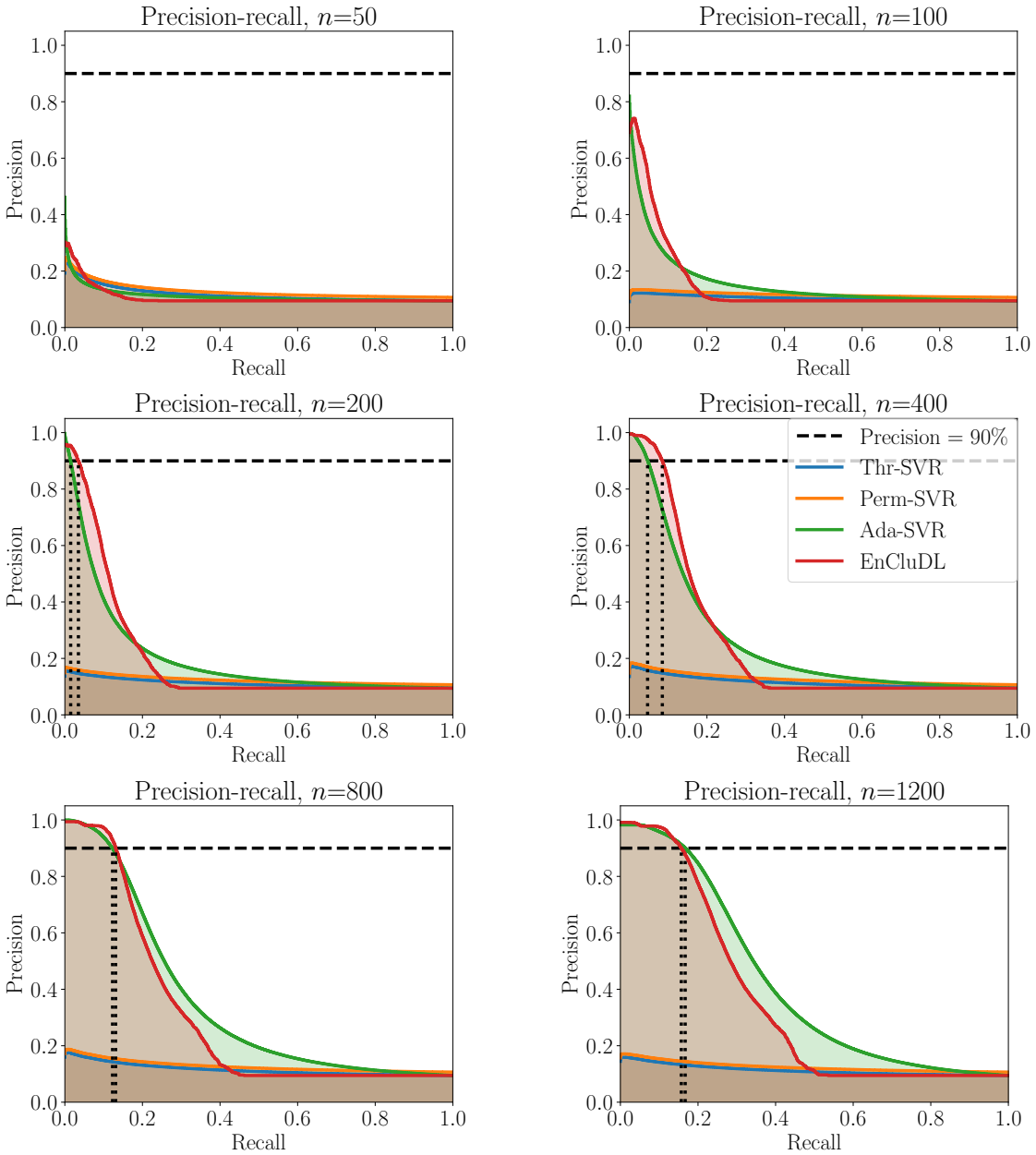


Figure 18: **Precision-recall curves on semi-simulated data with binary response vector.** The results of the experiment described in Sec. 4.2 with binary response show that Ada-SVR and EncluDL outperform alternatives in terms of feature recovery. These results are quite similar to the one presented in Fig. 6.

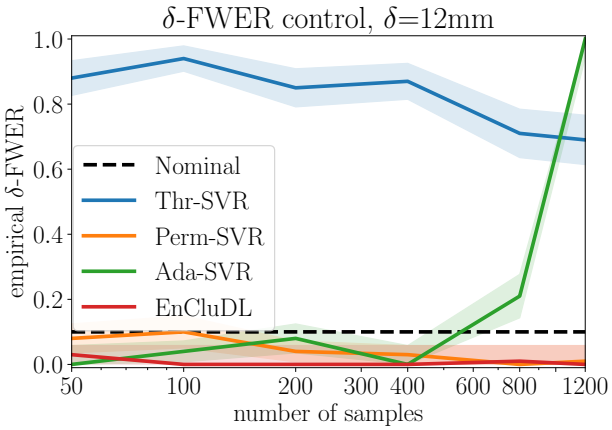


Figure 19: ***δ*-FWER control on semi-simulated data with binary response vector.** The results of the experiment described in Sec. 4.2 with binary response show that only Perm-SVR and EnCluDL actually control the .

δ-FWER.

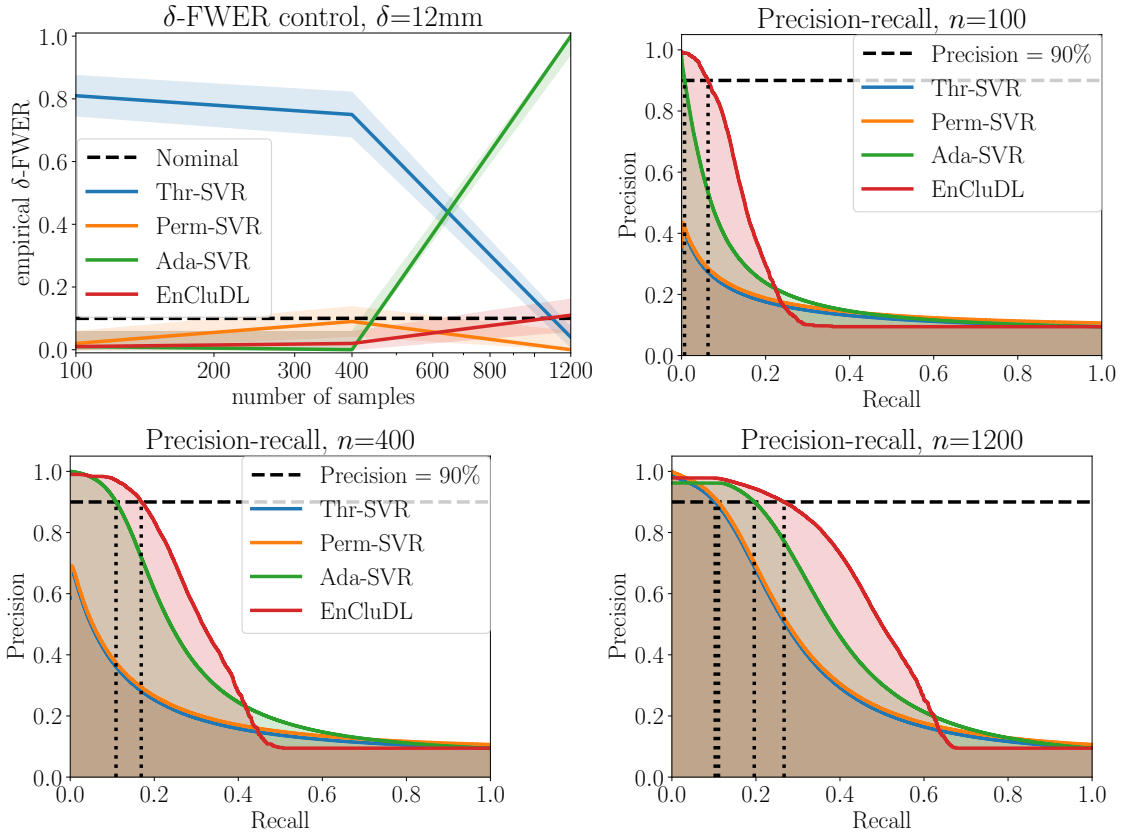


Figure 20: **δ-FWER control and precision-recall curves on semi-simulated data with continuous response vector with Laplace noise.** The results of the experiment described in Sec. 4.2 with Laplace noise are similar to the one presented in Fig. 6 for Gaussian noise.

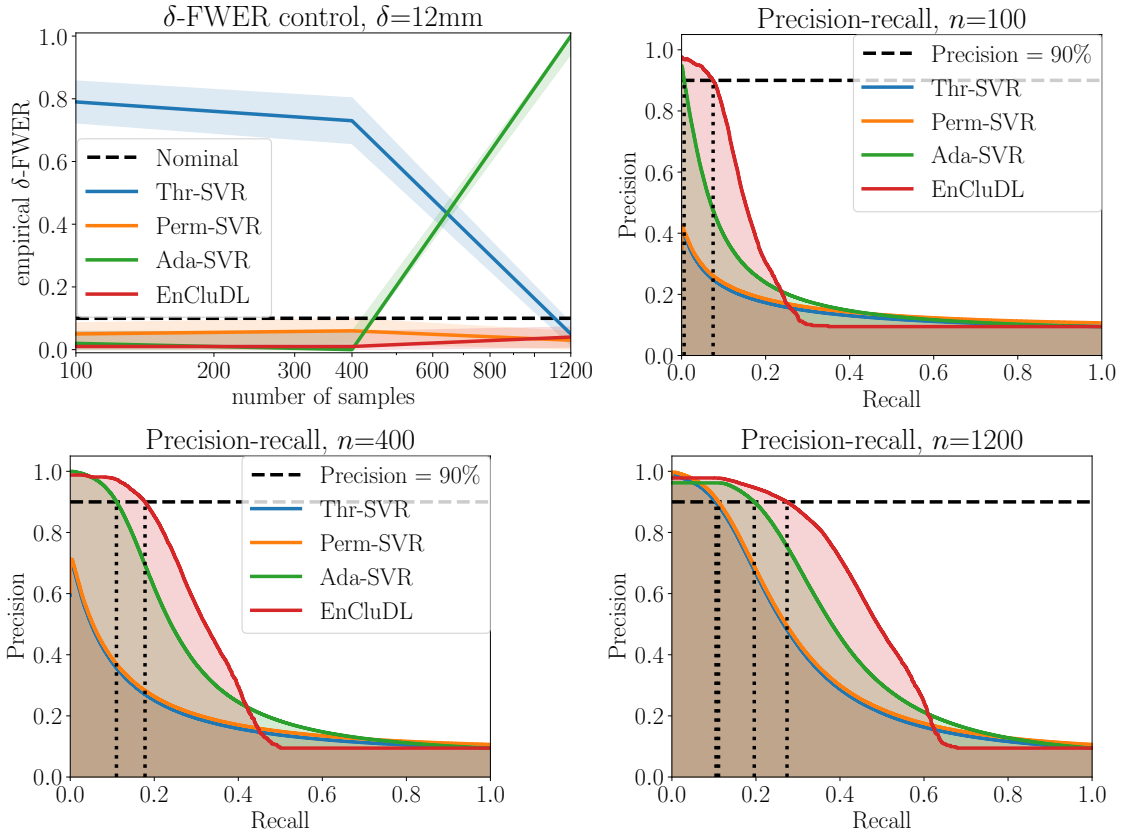


Figure 21: **δ-FWER control and precision-recall curves on semi-simulated data with continuous response vector with Student noise.** The results of the experiment described in Sec. 4.2 with Student (with 5 degrees of freedom) noise are similar to the one presented in Fig. 6 for Gaussian noise.

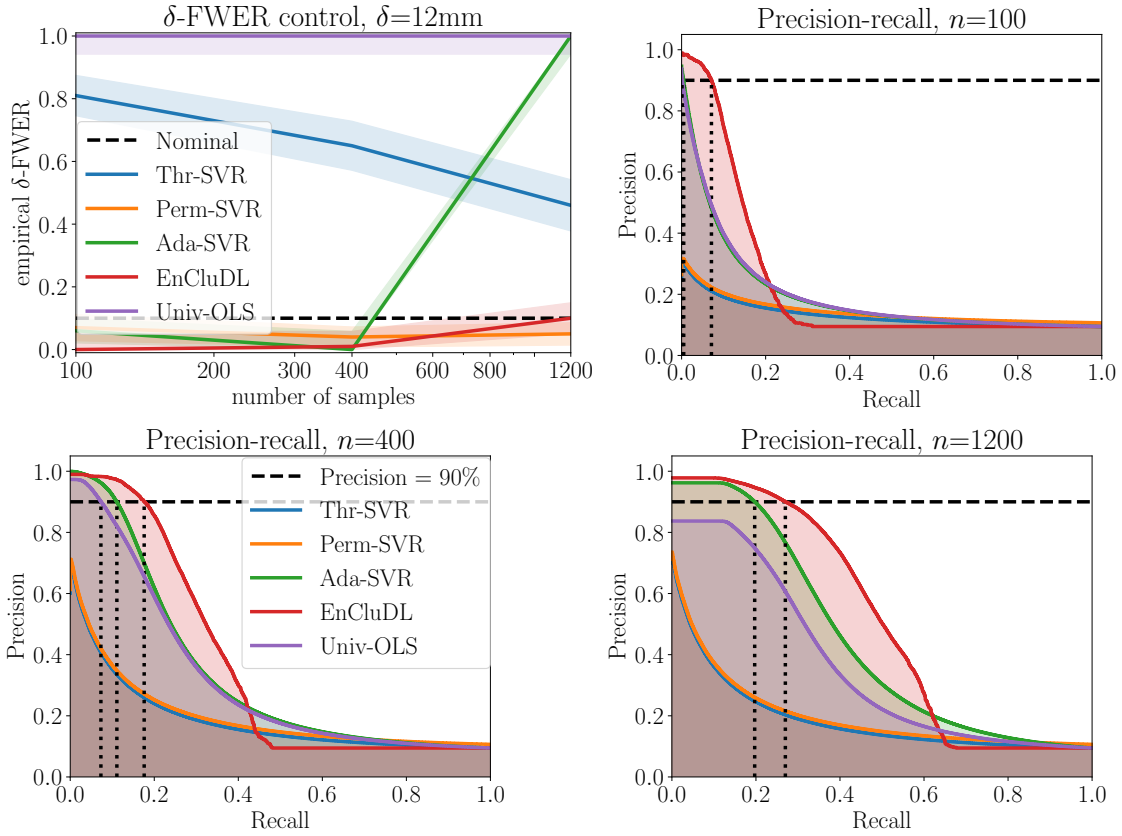
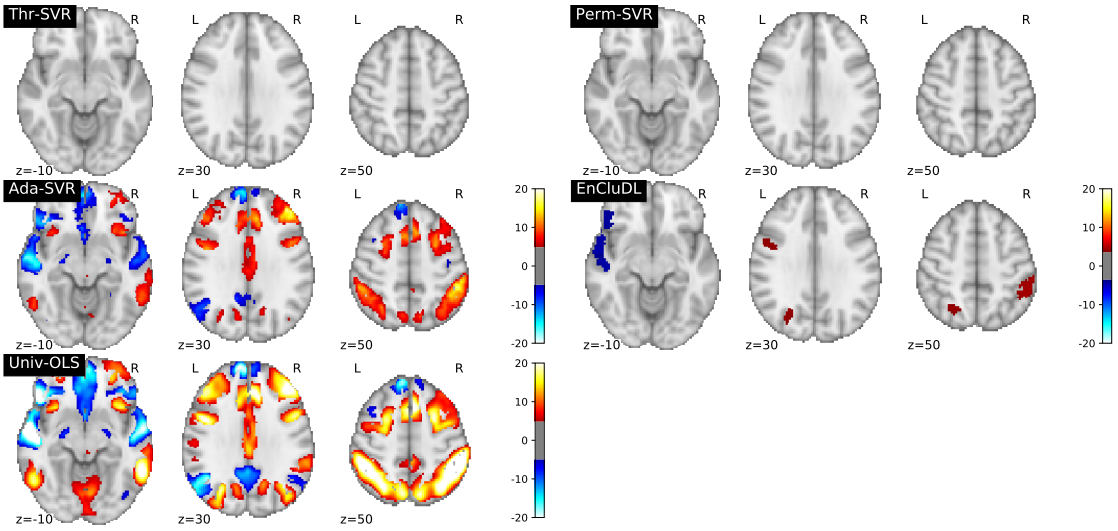
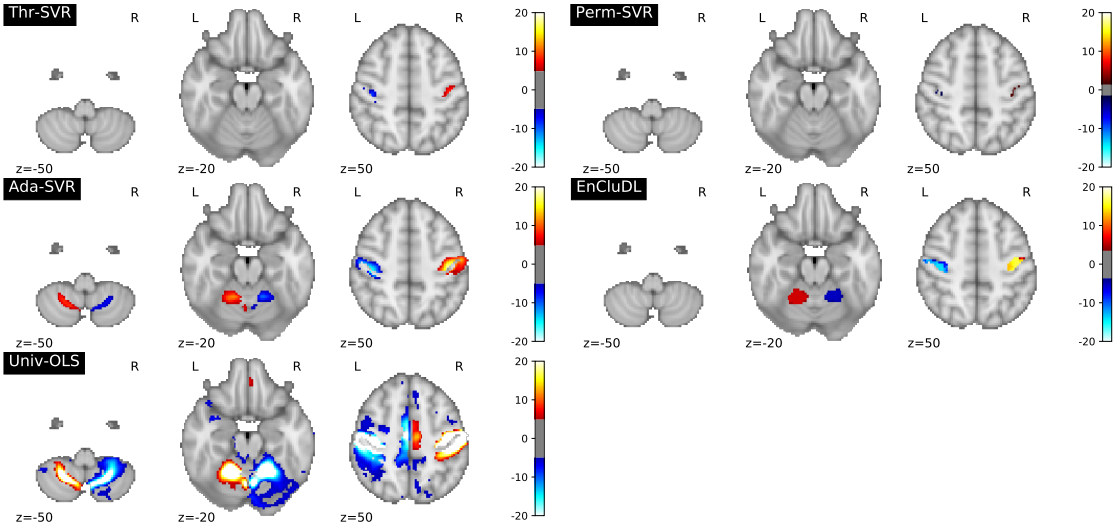


Figure 22: **δ-FWER control and precision-recall curves on semi-simulated data with continuous response vector including a univariate method.** These results show that the FWER control guaranteed by Univ-OLS for univariate inference does not match the control granted by EncluDL in the conditional paradigm. This is due the fact that the null hypotheses being tested are not the same.

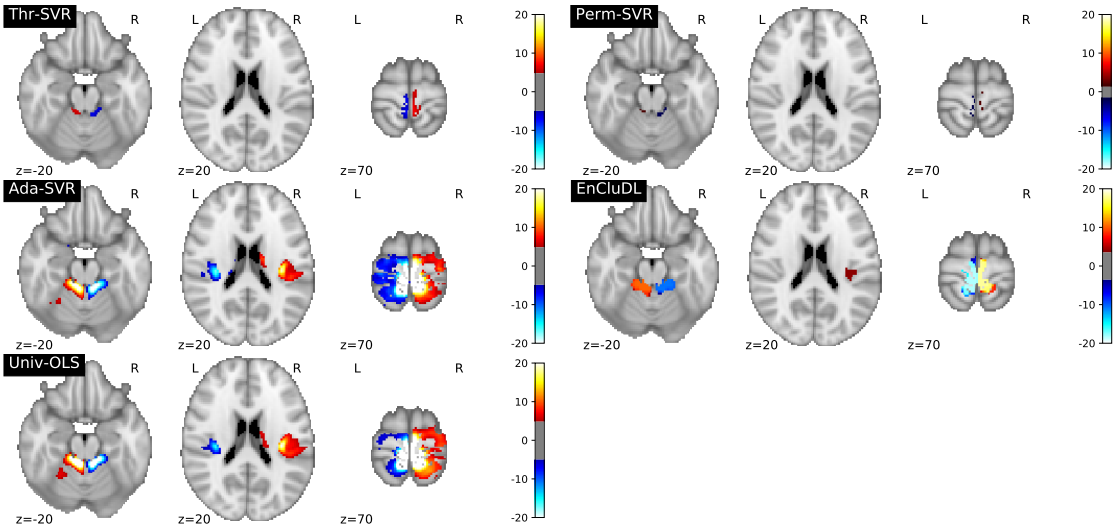


(c) Language

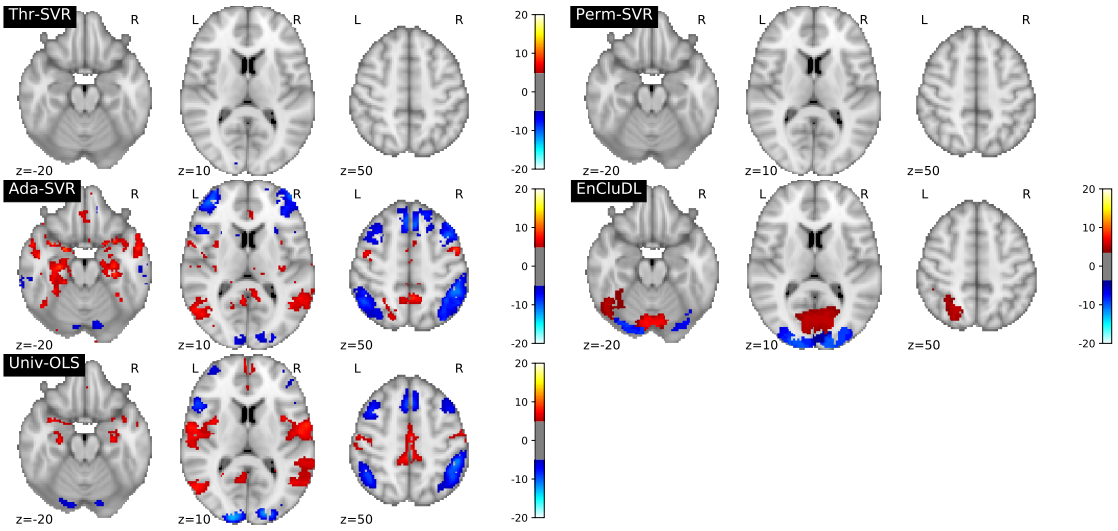


(d) Motor Hand

Figure 18a: cf. Fig. 23 for description.

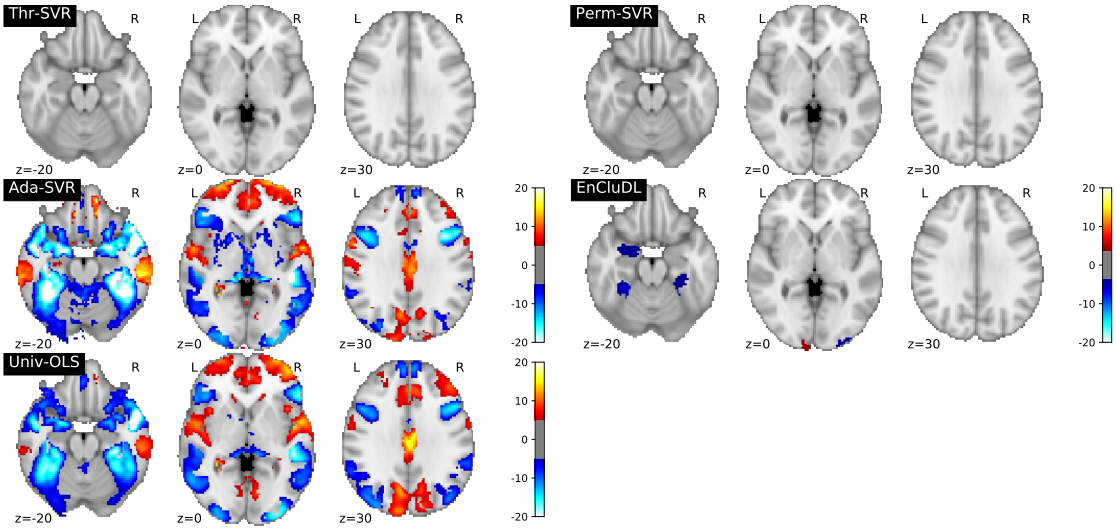


(e) Motor Foot



(f) Relational

Figure 18b: cf. Fig. 23 for description.



(g) Social

Figure 23: **Estimated predictive patterns on standard task fMRI dataset.** Here, we plot the results for five tasks of the experiment described in Sec. 4.6 thresholding the statistical maps such that the δ-FWER stays lower than 10% for $\delta = 12\,\text{mm}$. Qualitatively, EnCluDL discovers the most plausible patterns, Ada-SVR often makes dubious discoveries, patterns are too wide and implausible, while the two other methods exhibit a very weak statistical power. As discussed before, Univ-OLS provides complementary results that display marginal associations between voxel signals and the target. The results of emotion and gambling tasks are available in Fig. 7.

7.12 Prediction performance

In this section, we give results on the prediction performance of the methods. In Fig. 24, we plot the results of the experiment described in Sec. 4.7. We notice that the classification error rate is almost the same for SVR (the weight map of Thr-SVR and Perm-SVR) and Ada-SVR, their prediction performance is slightly better than the one of EnCluDL. Hence, we do not recommend using EncluDL to achieve state-of-the art prediction accuracy, but only for statistical inference purpose.

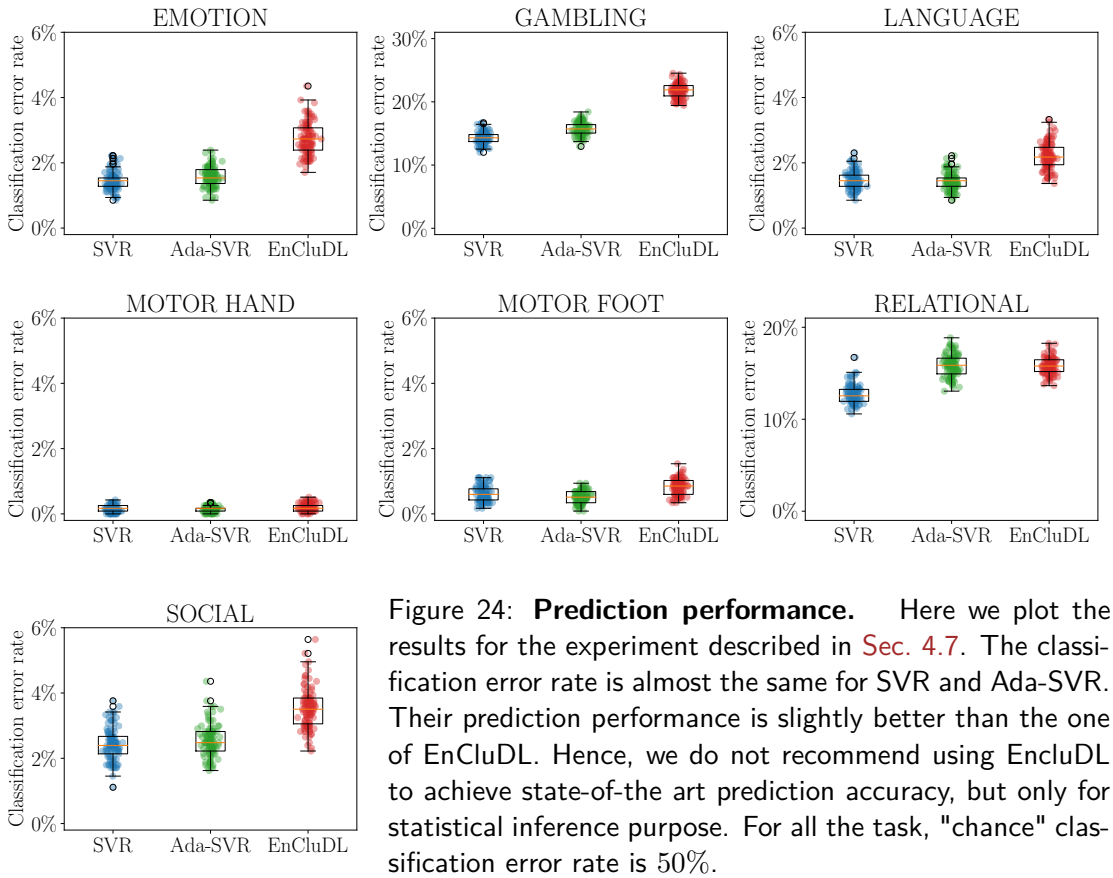


Figure 24: **Prediction performance.** Here we plot the results for the experiment described in Sec. 4.7. The classification error rate is almost the same for SVR and Ada-SVR. Their prediction performance is slightly better than the one of EnCluDL. Hence, we do not recommend using EncluDL to achieve state-of-the art prediction accuracy, but only for statistical inference purpose. For all the task, "chance" classification error rate is 50%.